Regulating AI: A Matrix for Gauging Impact and its Legal Implementation

by **Maurice Chiodo, Henning Grosse Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera, Herbert Zech ***

Abstract: This paper presents new and alternative avenues for regulating AI, understood in the broad sense of state intervention that directly or indirectly affects how AI is developed and deployed. In doing so, it also identifies two new regulatory axes to be addressed - Big/Small AI Impact, and Big/Small AI Teams - and the challenges associated with designing tools that are effective and tailored to these axes. Against this background, our proposals in this paper thus fall into two broad categories:

1. Forms of strict liability, as a means to incentivise operators of Small AI systems (where direct regulation is unlikely to be effective, and often unnecessary) to take steps preventing or at least minimising harm resulting from their AI.

2. A mechanism of constant monitoring through random sampling with human testing and determination, as well as associated cumulative "penalty points", to control the impact of Big AI systems on individuals and society before they cause large-scale, and in particular systematic, harm to society.

Our proposals aim to have the maximal possible impact on the behaviour of AI operators and on where and how AI is used, so as to prevent harm and encourage responsible development and use of AI. In our view, these proposals are implementable, policeable, enforceable, cost-effective, and technologically-neutral; they do not focus on what the AI is, but rather what it does. We consider different legal tools and mechanisms for implementing our proposals, and conclude that while elements of existing tort liability regimes, platform regulation as well as the EU AI Act can partially be relied on, law makers should consider adopting more tailored rules on both liability and constant monitoring.

Keywords: AI regulation, tort liability, EU AI Act, AI Oversight, Digital Services Act

© 2026 Maurice Chiodo, Henning Grosse Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera, Herbert Zech

Everybody may disseminate this article by electronic means and make it available for download under the terms and conditions of the Digital Peer Publishing Licence (DPPL). A copy of the license text may be obtained at http://nbn-resolving.de/urn:nbn:de:0009-dppl-v3-en8.

Recommended citation: Maurice Chiodo, Henning Grosse, Ruse-Khan, Dennis Müller, Lea Ossmann-Magiera, Herbert Zech, Regulating AI: A Matrix for Gauging Impact and its Legal Implementation, 17 (2026) JIPITEC 97 para 1.

A. Overview of the Problem

I. The Challenge of Regulating AI

1 It is well understood that the control and regulation of artificial intelligence[1] (AI) is both an important

* Authors are listed alphabetically by surname.

Maurice Chiodo: Centre for the Study of Existential Risk, University of Cambridge, UK. mcc56@cam.ac.uk

Henning Grosse Ruse-Khan: Faculty of Law, University of Cambridge, UK. hmg35@cam.ac.uk

Dennis Müller: Institute of Mathematics Education, University of Cologne, Cologne, Germany. dennis.mueller@uni-koeln.de

Lea Ossmann-Magiera: YPOG, Weizenbaum Institute for the Networked Society, Berlin, Germany. lea.ossmann-magiera@ypog.law

Herbert Zech: Humboldt-University of Berlin, Weizenbaum Institute for the Networked Society, Berlin, Germany. herbert.zech@hu-berlin.de

The authors wish to thank the attendees of the Weizenbaum Institute event in November 2023 and those of the International Law and Technology Governance class at NYU in March 2024 for their comments and suggestions, in particular Rochelle Dreyfuss. All errors remain the sole responsibility of the authors.

1 Understood as the turn from pre-programmed software

matter to address, and one for which no satisfactory mechanism exists yet. This is unsurprising, given how complicated such systems are. Indeed, an early 2024 editorial in Nature[2] mentions the EU AI Act has "gaps" that to resolve requires input from the scientific community. Due to "implementation challenges" (like a lack of harmonised standards for the AI Act's high-risk requirements) the European Commission also proposed to postpone the application of the AI Act (which was originally scheduled for 2 August 2026), especially for high-risk AI systems.[3]

2 Lawmaking is often reactive, occurring after a problem(s) arises in society - with an implicit expectation that new problems only manifest a few times before the law is adapted and catches up. However, current AI development is happening at an increasingly rapid pace and scale, such that ongoing attempts to adjust existing laws cannot keep up. Given the AI risks we discuss below, we suggest regulatory approaches are needed that can curtail future harms posed by AI *before their manifestations, often in the form of irreversible damage to individuals, groups and societies.*

3 The following examples illustrate these difficulties, showing the ever-increasing diversity of application domains, types of AI algorithms, and the multitude of ethical and legal challenges involved. These examples also foreshadow our suggestions to regulate AI by *team size* and *impact size* (in both cases counting the number of people involved/affected).

An example of an AI with a big impact, run by a small team:

4 Stability.ai is an AI text-to-image generating platform that produces images according to prompts from the user. It has been accused by Getty Images of significant copyright violations, allegedly using millions of Getty's images for AI training without adequate permission, and has been sued by Getty for over $1.8 trillion in damages.[4] While the English High Court rejected most of Getty's claims in its first instance decision in November 2025,[5] Getty subsequently won the right to appeal in the UK.[6] And as well as potentially infringing on the IP rights of countless artists,[7] Stability.ai (and similar generative AI tools) also run the risk of sending many creators and businesses in the creative industries out of business, by fundamentally shifting the process of visual content creation. Stability.ai impacted a "big" number of people in a relatively short timespan,[8] and

to machine learning. For a more rigorous definition, see the OECD definition in Russell et al., *Updates to the OECD's definition of an AI system explained*, OECD (Nov. 29, 2023), https://oecd.ai/en/wonk/ai-system-definition-update (last accessed Feb. 29, 2024).

2 *There are holes in Europe's AI Act — and researchers can help to fill them,* Editorial, Nature **625**, 216 (Jan. 10, 2024), https://www.nature.com/articles/d41586-024-00029-4 (last accessed Feb. 29, 2024).

3 Proposal for a Regulation of the European Parliament and of the Council amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI), COM(2025) 836 final, art. 1 sec. 31 lit. a. The reasons for the proposal are described in the explanatory memorandum, p. 1-3 of the proposal.

4 See Growcoot, *Getty Images is Suing Stable Diffusion for a Staggering $1.8 Trillion,* PetaPixel (Feb. 7, 2023), https://petapixel.com/2023/02/07/getty-images-are-suing-stable-diffusion-for-a-staggering-1-8-trillion/ (last accessed Feb. 29, 2024). Next to the lawsuit in the UK discussed below, Getty is also pursuing claims against Stability.ai in U.S. courts which, at the time of writing, are ongoing - see Getty's filings at the District Court of Northern California at https://www.courtlistener.com/docket/71112094/getty-images-us-inc-v-stability-ai-ltd/. See also the legal filing at GMG Editorial, *Getty Lawsuit Against Stable Diffusion*, Scribd (Mar. 2, 2023), https://www.scribd.com/document/624165761/Getty-Lawsuit-Against-Stable-Diffusion (last accessed Jan. 4, 2026).

5 *Getty Images v Stability AI,* [2025] EWHC 2863 (Ch), 4 November 2025. Already during the proceedings, Getty had to drop its principal claims for copyright infringement as none of the relevant copying activities related to training Stability.ai's Stable Diffusion model could be shown to have taken place within the jurisdiction of English courts. And while Getty succeeded with some elements of its trademark infringement claims (based on the memorisation and reproduction of the Getty Image logo by output generated when using Stability.ai's model), the most relevant remaining copyright claim (based on the argument that the image generation service offered by Stability.ai's Stable Diffusion model constituted an infringing copy imported into the UK under section 22 & 23 CDPA) has been rejected.

6 The Court of Appeal allowed Getty to challenge the High Court's finding on whether the Stable Diffusion AI model can be considered an infringing copy - see Barker, *High Court grants Getty permission to appeal copyright aspects of GenAI claim* (Dec. 19, 2025) - online at: https://thelens.slaughterandmay.com/post/102lydj/high-court-grants-getty-permission-to-appeal-copyright-aspects-of-gen-ai-claim (last accessed Jan. 5, 2026).

7 For further discussion about generative AI and copyright law, see for example Zirpoli, *Generative Artificial Intelligence and Copyright Law*, Congressional Research Service (Feb. 24, 2023), https://digitalcommons.unl.edu/scholcom/243 (last accessed Feb. 29, 2024).

8 Stability.ai, https://stability.ai/about (last accessed Feb. 29, 2024).

moreover was a small company, that at the time had few employees.[9]

An example of an AI with a small impact, run by a small team:

5 Clearview AI is a company that produces AI-powered facial recognition software. They claim to have scraped the Internet, including social networking sites, for over 10 billion photos of individuals,[10] which was then used in the production of their software. Clearview AI have faced numerous lawsuits and fines relating to privacy violations from countries around the world.[11] The company carried out its work in relative obscurity for 3 years, until the New York Times wrote about them in 2020.[12] As they were a small organisation, and effectively impacted only a small number of people, they were able to evade public scrutiny, and even public *awareness*, for a long time.

An example of an AI with a big impact, run by a big team:

6 Large Recommendation Systems, such as those run by Amazon, are (often AI-powered) tools that recommend items to consumers on large platforms.[13] These can be items for purchase (in the case of Amazon), media to view (in the case of Netflix, YouTube), etc. Such systems face a myriad of ethical issues and potential harms, such as inappropriate content, privacy, autonomy and personal identity, opacity, fairness, and wide scale social effects.[14] Many of these platforms employ tens of thousands of employees,[15] or even millions.[16] And the impact of these recommenders is massive: YouTube and Amazon have billions of users,[17] and billions in sales,[18] respectively. The sheer scale of these AI-based systems being deployed can have a significant impact on societies.

An example of an AI with a small impact, run by a big team:

7 In 2022, a team of researchers at a drug discovery company were developing AI to postulate new chemical compounds with the minimum possible toxicity, which could then be synthesised, tested, and eventually applied to medical uses. However, by simply "reversing" the learning objectives to instead postulate new chemical compounds with the *maximum* possible toxicity, the team developed an AI that identified "many other known chemical warfare agents", as well as "new molecules [that] were predicted to be more toxic … than publicly known chemical warfare agents".[19] The impact space of such developments would, though severe for affected individuals, potentially only be on a small proportion of the population. But it would not be an easy process. Those who produced the research had extensive experience and expertise, having "spent decades using computers and AI to improve human health".[20] Moreover, the testing, verification, and production of such postulated compounds requires significant understanding of chemistry, as well as elaborate and expensive production facilities; to do all this requires a large, well-resourced organisation.

8 Given the myriad of possible use cases and scenarios, how can lawmakers foresee all possible uses and harms of AI, so as to legislate, regulate, or otherwise intervene to prevent or address them? This might not be necessary. In this paper, we present a combination of disincentives for AI operators, and

9 195 according to their LinkedIn page: Stability.ai, LinkedIn, https://uk.linkedin.com/company/stability-ai (last accessed Dec. 6, 2023).

10 See Knight, *Clearview AI Has New Tools to Identify You in Photos*, Wired (Oct. 4, 2021), https://www.wired.com/story/clearview-ai-new-tools-identify-you-photos/ (last accessed Feb. 29, 2024).

11 See Mac, Hill, *Clearview AI settles suit and agrees to limit sales of facial recognition database,* New York Times (May 9, 2022), https://www.nytimes.com/2022/05/09/technology/clearview-ai-suit.html (last accessed Feb. 29, 2024).

12 See Hill, *The Secretive Company That Might End Privacy as We Know It*, The New York Times (Jan. 18, 2020), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html (last accessed Feb. 29, 2024).

13 See Milano et al., *Recommender systems and their ethical challenges*, 35 AI & SOCIETY 957 (2020), https://link.springer.com/article/10.1007/s00146-020-00950-y (last accessed Feb. 29, 2024).

14 ibid.

15 See Andreeva, *Netflix Hit By Layoffs; About 150 Mostly U.S.-Based Employees Affected*, Deadline (May 17, 2022), https://deadline.com/2022/05/netflix-layoffs-us-employees-1235026085/ (last accessed Feb. 29, 2024).

16 See Yahoo Finance, Amazon Inc., https://finance.yahoo.com/quote/AMZN/profile (last accessed Feb. 29, 2024).

17 See Wong, Bottorff, *Top Social Media Statistics And Trends Of 2024*, Forbes, https://www.forbes.com/advisor/business/social-media-statistics/ (last accessed Feb. 29, 2024).

18 See Statista, *Net e-commerce sales of Amazon from 2018 to 2027, by business model*, https://www.statista.com/statistics/1309709/amazon-e-commerce-retail-sales-business-models/ (last accessed Feb. 29, 2024).

19 See Urbina et al., *Dual use of artificial-intelligence-powered drug discovery*, 4 Nature Machine Intelligence 189 (2022), https://www.nature.com/articles/s42256-022-00465-9 (last accessed Feb. 29, 2024).

20 ibid.

monitoring of AI behaviour, which we believe may significantly prevent harm. After all, we believe that the primary reason for using legal tools to regulate AI lies in the prevention of harm.

II. Difficulties with Existing AI Regulatory Proposals and Ideas

9 Many suggestions for ways to regulate and control AI can be found in the academic literature,[21] draft laws, and regulatory and legal practice,[22] even though the current narrative of a US-China "race" to Artificial General Intelligence (AGI) has "been used to justify reduced regulatory oversight".[23] However, as we outline below, most of them are infeasible or leave substantial gaps in the detection or mitigation of harm. For example:

1) Proposals to introduce algorithms to monitor algorithms,[24] or go further and challenge algorithms such as a "Public AI",[25] might seem tempting; deploying a technological solution to an apparently technological problem. But these suggestions contain hurdles. Firstly, verifying that the good AI, which checks (or challenges) the potentially bad AI, is doing an effective job is very difficult; do we not also need an AI to check the checker, and so on? If building such a tool were easy, it could have been incorporated into the original AI. Secondly, the cost, time, and expertise required to produce such an "AI checker" could be prohibitive. Existing AI systems, such as ChatGPT, require access to massive training data sets, massive compute power, large teams of extremely expensive experts,[26] consume massive amounts of electricity and other resources,[27] and take a long time to set up.[28] Moreover, given that these checkers would be needed for a particular context, jurisdiction, or social setting, different states and/or regulatory entities would be required to build and maintain their own versions. Sharing one general version is unlikely to be satisfactory. So only the most wealthy, well-resourced entities or states would

21 For an overview of existing guidelines and related proposals see Schiff et al., *What's Next for AI Ethics, Policy, and Governance? A Global Overview, in* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20, Feb. 7-8, 2020), pages 153-158, https://doi.org/10.1145/3375627.3375804 (last accessed Feb. 29, 2024); Hagendorff, *The ethics of AI ethics: An evaluation of guidelines*, 30 Minds and Machines 99 (2020), https://doi.org/10.1007/s11023-020-09517-8 (last accessed Feb. 29, 2024); Ryan, Stahl, *Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications*, 19(1) Communication and Ethics in Society 61 (2021), http://dx.doi.org/10.1108/JICES-12-2019-0138 (last accessed Feb. 29, 2024).

22 For an overview on some recent regulatory and legal approaches and a comparison between the AI Act and UK's legislative approach see Hacker, *AI Regulation in Europe: From the AI Act to Future Regulatory Challenges*, arxiv (Oct. 6, 2023), https://arxiv.org/abs/2310.04072 (last accessed Feb. 29, 2024).

23 Ó hÉigeartaigh, *The most dangerous fiction: The rhetoric and reality of the AI race.* SSRN Preprint (2025) http://dx.doi.org/10.2139/ssrn.5278644 (last accessed Jan. 4, 2026).

24 See Machin et al., *Specifying Safety Monitors for Autonomous Systems Using Model-Checking, in* Proceedings of the International Conference on Computer Safety, Reliability and Security (SAFECOMP, Sep. 2014), pages 262-277, https://link.springer.com/chapter/10.1007/978-3-319-10506-2_18 (last accessed Feb. 29, 2024).

25 See Elkin-Koren, *Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence*, 7(2) Big Data & Society (Jul. 29, 2020), https://journals.sagepub.com/doi/full/10.1177/2053951720932296 (last accessed Feb. 29, 2024).

26 Butcher, *A.I. Specialists Can Earn as Much as NFL Players*, Dice (Jul. 17, 2017), https://www.dice.com/career-advice/a-i-specialists-earn-much-nfl-players (last accessed Feb. 29, 2024).

27 Freitag et al., *The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations*, 2(9) Patterns 1 (2021), https://doi.org/10.1016/j.patter.2021.100340 (last accessed Feb. 29, 2024); Hacker, *AI Regulation in Europe: From the AI Act to Future Regulatory Challenges*, arxiv (Oct. 6, 2023), https://arxiv.org/abs/2310.04072 (last accessed Feb. 29, 2024); Saul, Bass, *Artificial Intelligence Is Booming - So Is Its Carbon Footprint*, Bloomberg (Mar. 9, 2023) https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure (last accessed Feb. 29, 2024).

28 cf. Marr, *A Short History Of ChatGPT: How We Got To Where We Are Today*, Forbes (May 19, 2023), https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today (last accessed Feb. 29, 2024).

stand a chance of producing such a tool, if at all.

2) An expectation of AI operators to self-regulate and produce responsible and safe AI, even without regulatory obligation, might initially seem reasonable. After all, existing liability regimes might already incentivise AI operators to avoid causing harm, as that may trigger obligations to undo the harm or to compensate (but see 4. in this list). However, sufficient self-regulation is unlikely to occur in practice, for many reasons. Firstly, there are no certified "licensing" requirements on those developing AI or other mathematical products; it is not a professional vocation, unlike law or medicine.[29] And while standards continue to be set in various ways,[30] the mechanisms to enforce adherence to such standards remain weak. Second, AI developers (working for AI operators) will likely come from various training backgrounds, including computer science, mathematics, physics, and engineering. These provide minimal compulsory training in AI ethics and responsible research and innovation,[31] so developers may have had little if any such training.[32] At the moment, responsible development and ethical training are not forced upon developers, neither by a professional oversight body, nor by their own intrinsic disciplinary culture. Finally, our experience working with startups and teaching ethics to mathematicians,[33] has led to encounters with technical experts who profess that they are "just doing the technological development, and that it is not their responsibility to ensure its safety or ethics". Just as the legal and regulatory community might expect operators to take it upon themselves to produce safe tools and products, the operators might look towards lawmakers for legislative guidance to establish at least the basic regulatory framework with an expectation that they only need to "follow the law". Thus, the responsibility to regulate AI is simultaneously passed back to lawmakers. Going beyond the existing legal framework by self-imposing further checks or constraints will also potentially affect profitability and hence concern investors or share-holders, making self-regulation unlikely unless all relevant actors in the industry agree.

3) The ability to regulate AI often presumes that its use is known, and its impact understood, or at least verifiable. However, relying on the possibility of detecting the presence or use of AI can be misplaced. Many uses of AI fail to leave any trace or "watermark". It is hence difficult if not impossible to tell when AI has been utilised and/or served as one, or even the essential, element or process leading to an outcome impacting on others. This is particularly true for AI systems that make choices, decisions, or recommendations. For example, AI used to shortlist job applicants for interviews based on their CV data may raise many ethical issues,[34] including that there may be no way of identifying the reliance on AI in the first place. The shortlisting could have been done by hand, or by some elementary algorithm giving a rudimentary score such as "(years of experience) divided by (salary requested)", or by an AI trained on a large dataset; all are plausible. Similarly, AI might be used as part of a larger techno-social system, including search engine recommendations, online advert distribution, calculating insurance premiums, and determining public transport timetables. In all these applications, and many others, the output is straightforward: the user or customer is presented with a list of search results, shown an advert, given an insurance cost, or handed a bus map and timetable. An end-user or impacted party, as much as a regulator, has almost no way of knowing if, or how, AI was used in such a process.

4) A more 'hands-off' regulatory approach might suggest relying on existing rules of tort liability (for example, for faulty products or substandard service provision) and apply them to the use of AI once harm has been caused. However, the assumption that this might induce liability-avoiding behaviour ex ante, and - more importantly - of being able to restitute or compensate for the relevant harm ex post-impact, may not always hold. Akin to the global financial crisis of 2008, when the world saw banks and insurance companies deemed "too big to fail", some AI processes may cause damages

29 Müller et al., *A Hippocratic Oath for Mathematicians? Mapping the Landscape of Ethics in Mathematics*, 28(5) Sci Eng Ethics 41, https://doi.org/10.1007/s11948-022-00389-y (last accessed Feb. 29, 2024).

30 Kilian et al., European *AI Standards - Technical Standardization and Implementation Challenges under the EU AI Act* (Feb. 26, 2025). SSRN Preprint (2025) https://dx.doi.org/10.2139/ssrn.5155591 (last accessed Jan. 5, 2026).

31 Horton et al., *Embedding Ethics in Computer Science Courses: Does it Work?, in* Proceedings of the 53rd ACM Technical Symposium on Computer Science Education (SIGCSE 2022, Mar. 3-5, 2022), pages 481-487, https://doi.org/10.1145/3478431.3499407 (last accessed Feb. 29, 2024).

32 Müller et al., *A Hippocratic Oath for Mathematicians? Mapping the Landscape of Ethics in Mathematics*, 28(5) Sci Eng Ethics 41, https://doi.org/10.1007/s11948-022-00389-y (last accessed Feb. 29, 2024).

33 Chiodo, *The Ethics in Mathematics Project*, Cambridge University Ethics in Mathematics Project (Dec. 5, 2025), https://www.ethics-in-mathematics.com (last accessed Dec. 5, 2025).

34 Kodiyan, *An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool* (Nov. 2019), https://www.researchgate.net/publication/337331539_An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazon%27s_AI_based_hiring_tool (last accessed Feb. 29, 2024).

which are too big to undo, compensate, or fine for and hence would have created harmful consequences that are too significant to 'fix' (i.e., restitute in ways commonly set out in tort law). Large-scale technology-induced harm, such as oil spills and other disasters, are often dealt with by enormous human effort and commensurate payment from the organisation at fault.[35] However, this assumes that the entity responsible, its insurer, or - as a matter of last resort - the relevant polity affected by the harm has sufficient resources to undo or otherwise address the harm caused. Given the scale and impact some uses of AI can have,[36] it might not necessarily be the case that the damage done can in fact be 'fixed' at all.[37] Since, as a matter of last resort, humanity at large, other living beings, and our planet shoulder those risks which cannot be restituted or compensated for if materialised, we believe that preventative measures are needed, not just post-harm recourse.

5) Conceptions of legal responsibility under private law generally provide an effective recourse for individuals who have been harmed or suffered losses as the result of the actions of another. As suggested above, it may be tempting to assume that tort and related liability regimes can operate satisfactorily to address the harms caused by AI. But, in addition to the ex post vs ex ante issue already considered, the entity producing or deploying the AI may in practice escape liability due to difficulties in establishing that AI has been involved, showing a sufficient causal link, or offering relevant evidence.[38] In addition, some of the most important actors involved in producing or deploying "Big AI" (as defined below) at scale (Microsoft, Google, Meta, etc.) are so well-resourced that an individual harmed stands almost no realistic chance of successfully pursuing claims against them in litigation. In addition to these hurdles, major further practical roadblocks are likely to arise in the form of finding a court with jurisdiction and determining the applicable law, contractual limits of liability imposed via terms of service (if valid), and the inability to enforce judgments in jurisdictions where the relevant actors have significant assets. Furthermore, existing laws might not always sufficiently conceptualise the end-user's interest in a way that captures the potential harm done by AI. If, for example, a large video sharing platform, such as YouTube, takes down a video of a birthday party on the basis of someone's copyright claim that is automatically enforced via Content ID,[39] most copyright laws will struggle to offer a legally recognised interest that the uploader can put forward in claiming reinstatement of the video on the platform. While users certainly enjoy (fair use and other) defences against infringement claims by rights holders, those are likely to remain ineffective when AI tools automatically enforce such claims across the platform: Not only is AI (so far) unable to comprehend contextual issues such as (fair) uses for the purpose of parody, critique or quotation - but conventional copyright doctrine is only beginning to respond to automated private ordering tools when deployed to prevent a defensible use, or to enforce the monetisation of such

35 Cf. Taylor, *Catastrophes and Insurance*, in Global Catastrophic Risks 164-183 (Bostrom, Cirkovic, Jul. 2008), https://doi.org/10.1093/oso/9780198570509.003.0012 (last accessed Feb. 29, 2024); Chang et al., *Consequences of oil spills: a review and framework for informing planning*, 19(2) Ecology and Society 26 (2014), http://dx.doi.org/10.5751/ES-06406-190226 (last accessed Feb. 27, 2024); Leach, *Litigating Catastrophe*, 110 California Law Review 2149 (2022), https://doi.org/10.15779/Z38MK65918 (last accessed Feb. 27, 2024).

36 Crawford, *Atlas of AI. Power, politics, and the planetary costs of artificial intelligence*, Yale University Press (Apr. 6, 2021).

38 Schütte et al., *Damages liability for harm caused by Artificial Intelligence–EU law*, 69 Helsinki Legal Studies Research Paper 1 (2021), https://dx.doi.org/10.2139/ssrn.3897839 (last accessed Feb. 27, 2024).

39 Grosse Ruse-Khan, *Global Content Protection Through Automation – A Transnational Law Perspective*, 49 International Review of Intellectual Property and Competition Law 1017 (2018), https://doi.org/10.17863/CAM.35223 (last accessed Feb. 28, 2024).

8,5 Pt

use in favour of someone claiming copyright.[40]

III. AI as a Product or as a Service?

10 In addition to these shortcomings in the recent proposals to regulate AI, there is an overarching ambiguity in terms of classifying AI within conventional categories of commerce: some AI systems may be embedded into a product or understood as constituting a (digital) product, while others may be seen as part of providing a service, whereas many might involve a combination of both. There is an ongoing debate about this distinction, in particular within the EU: The EU Artificial Intelligence Act (AI Act) is conceived as a product safety instrument or product regulation. The European legislator thus treats AI systems as products. This is due to the fact that product regulation is harmonised and well-functioning since the EU's New Legislative Framework (NLF) entered into force. The NLF provides for a coherent certification and conformity assessment procedure and a uniform market surveillance for products placed on the market in the EU. The provisions of the AI Act are to be integrated into these legal structures and existing processes and procedures. In particular, the interplay between abstract legal requirements from legislative acts and the concrete instructions for actions from technical norms and standards is to be used for AI systems. In concrete terms, the method of co-regulation, which has established itself as fast, flexible and practicable and in which the EU legislator lays down basic principles while standardisation bodies determine the concrete action to be taken, is to be applied to AI systems.

11 However, certain AI systems (even those embedded in a product) have very different characteristics from conventional products. Some AI systems do for example change constantly after being placed on the market, especially if the training phase is not finished at the time of market entry, or if the AI is constantly being revised and improved. For some AI systems it is impossible to foresee what they will be used for at a later stage (e.g., General Purpose AI - GPAI) which makes an ex ante conformity assessment difficult to impossible.

12 Moreover, AI systems are often offered as a service ("AIaaS") and not as a product. This may be due to the operators' efforts to maintain trade secrecy and because producing and deploying AI systems requires an enormous computing capacity. By using AIaaS, smaller companies gain access to cutting-edge and up-to-date AI developments which they would not be able to use otherwise, since they may be missing the necessary (training) data, know-how, and computing resources to create, test and maintain these AI systems on their own.[41]

13 In contrast to products, regulating services - while not uncommon in regards to financial, insurance or legal industries as well as platforms - is a significantly more complex exercise, not least since there is no specific tangible item for the regulator to focus on. Regulating AIaaS directly does not seem to be on the regulatory agenda at the time of writing. When comparing the two concepts, however, it shows that the extent to which the actors involved can actually exercise influence varies greatly depending on whether an AI system is placed on the market as a product or the AI is offered as a service, which is why the legal responsibilities should be distributed accordingly. While AIaaS providers have a much bigger (or exclusive) and ongoing influence and impact on the training of their systems as well as the use, those same providers of AI lose that influence and often their ability to control their AI after placing such an AI-equipped product on the market. This may happen if, for example, a deployer makes, or is able to make, substantial changes to the AI.

IV. Terminology Concerning the Actors Involved

14 Regardless of the distinction between product and service, there is a distinction between different actors involved in the various stages of the lifecycle of AI systems. Based on the general distinction in product regulation, we differentiate between the producer and the user side. Following the terminology of the AI Act, we call the actors on the producer side *providers* and actors on the user side *deployers* (clarifying that end users are not implied herewith).

15 Both, providers and deployers, can be subsumed under a broader understanding of *operators*, art. 3 sec. 8 AI Act.[42] However, the term operators may also be understood in a narrower sense, relating to users responsible for the safety of products, art. 3

40 Grosse Ruse-Khan, *Automated Copyright Enforcement Online: From Blocking to Monetization of User-Generated Content* (Mar. 1, 2020), https://dx.doi.org/10.2139/ssrn.3565071 (last accessed Feb. 29, 2024).

41 Cobbe, Singh, *Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges*, 42 Computer Law & Security Review 105573, 2 (2021), https://doi.org/10.1016/j.clsr.2021.105573 (last accessed Feb. 28, 2024).

42 Regulation of the European Parliament and of the Council 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), hereinafter AI Act.

lit. e of the European Parliament's (now withdrawn) proposal for the introduction of an AI liability (see C.I.2).[43]

16 Throughout this paper, we use the terms *providers* on the producer side and *deployers* on the application side; the term *operator* is used as a term encompassing both. We will occasionally use the term *developers* in specific reference to those (individuals or teams) who are doing AI work within an AI operator. We use the term *users* to describe those people who engage with, or are impacted by, AI systems. These are summarised in Figure 1.

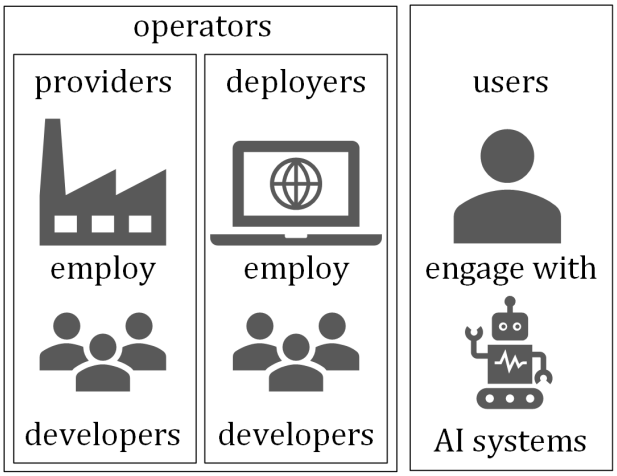

Figure 1: Terminology and actors involved in AI

V. A Feasible Way of Classifying AI

17 In order to adequately regulate AI, one needs to differentiate it along many axes. While the risk axis used by the EU in the AI Act[44] is valuable, it heavily depends on understanding, in advance, the growing list of application domains and types of AI systems, and on the capacity for classifying AI risks accordingly. Therefore, rather than building on a (potentially unlimited) number of different application domains, we propose to open two *additional* axes in addition to the existing EU risk classification, and further distinguish AI by its:

- Impact size, described as **Small AI** and **Big AI**, and
- Team size, described as **Small Team** and **Big Team**.

18 This leads to the following 2x2 matrix (i.e., grid) of "types" of AI, which we explain below.

	Big AI (big user base, big social impact)	**Small AI (small user base, small "direct" social impact, possible larger "indirect" impact)**
Big team	Examples: Twitter, Facebook, Google, Garmin, large retailers ● AI is usually part of a larger range of products or services. The product can be a corporeal item into which AI is integrated (e.g., in smartwatches). ● It takes many years to become a Big Team, working on Big AI. ● Many entities in this corner can also be called "gatekeepers".	Examples: manufacturers of specialised products, smaller medical or pharmaceutical companies ● AI is often part of an expensive product that is used by a small, often specialised group of people or sold to other businesses (B2B). ● AI is used to create and manufacture other tools or items. ● These companies are occasionally called "hidden champions" in the business literature: due to their unique role, they may be invisible to the non-expert or to people working in different areas, despite their size.
Small team	Examples: (early) OpenAI with ChatGPT, (early) Stability.ai with stable diffusion ● AI is often the only product, it gets sold to other companies or goes viral. ● The transition from(Small Team, Small AI) to (Small Team, Big AI) can be extremely fast, e.g., ChatGPT had 100million users within 2 months of launch.	Examples: Most startups in their early stages ● AI systems that stay in this corner often solve a niche problem, and regularly come with very specialised use-cases. ● For users and non-domain specialists it can be hard to understand who the AI providers are, and where the AI is used or sold. ● Startups are in this corner before they get big.

Figure 2: Matrix of AI types: Small AI vs Big AI, and Small Team vs Big Team

19 We do not specify precisely the impact size of "Big AI", in part because "impact" is a very difficult term to define. However, one useful proxy for impact, especially when considering risk and the potential for harm (as we do in this paper), is the *number of people (directly) affected.*[45] This is not a perfect measure, but nonetheless a useful measure of "impact". In addition, we note that the EU Digital Services Act (DSA) defines very large online platforms and very large online search engines (VLOPs and VLOSEs, together VLOPSEs) to be those which impact over 10% of the EU population.[46] Thus, taking "Big AI" to

43 Zech, *The European Parliament proposal for a regulation on liability for the operation of artificial intelligence systems,* 14 Zeitschrift für Geistiges Eigentum/Intellectual Property Journal 293, 298 (2022), https://doi.org/10.1628/zge-2022-0022 (last accessed Feb. 27, 2024).

44 See the "risk pyramid" in European Commission, *Shaping Europe's digital future: A European approach to artificial intelligence*, European Union (Jan. 31, 2024), https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (last accessed Feb. 28, 2024).

45 Which can come in many forms, such as "number of users", "number of people the AI makes decisions about", "number of customers sent AI-generated content", etc.

46 Art. 33 sec. 2 Regulation of the European Parliament and of the Council 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 1, hereinafter: DSA. See also section C.IV for a discussion on the DSA as an implementation framework for our proposals.

mean "AI that impacts over 10% of the population in question" would be a reasonable starting point for our discussion, though a more rigorous definition would be needed for any legislation.

20 Similarly, we do not specify precisely the size of a "Big Team". We envisage that any company which is publicly listed (or meets the threshold for public listing) would come under such a category. Or any company that meets a minimum number of employees (of, say, 1000). As such thresholds are only set for the purpose of our analysis, there is a degree of flexibility in how to interpret this term. Moreover, in our analysis we assume that "team size" correlates to the size of the overall company or organisation creating the AI, rather than the number of people directly involved in the development work. This is because all the factors that we are concerned with in Big Teams (internalised incentive to avoid doing harm, incentive for the company to not go bust, ability to defend against lawsuits, etc.) come from the overall size of the company.

1. Small Team - Small AI

21 The number of graduates focusing their studies on AI is rapidly rising and there is an ever-increasing number of companies creating AI tools.[47] Many of these companies with <50(ish) employees, in particular many startups, are producing AI tools, products and services for use in the open market. They regularly act as B2B suppliers, creating tools that other, often more prominent, corporations use for their day-to-day operations or as part of a larger product. These companies, and the work that they do, are often "invisible". Few people know they exist, and even those impacted are often unaware this is (partially) as a result of AI. It is also in this sense that Small AI leaves no easily detectable trace or "watermark". Such an AI system may still (in)directly impact millions of people, but not normally very large groups on a global scale (i.e., billions of people), and so can remain under the legal and regulatory radar for a long time. Their impact (as an individual organisation) is "small" when compared to big AI operators, and in comparison to the proportion of the overall population affected by their AI. As discussed in Section C.IV.2, digital regulation in the EU includes specific allowances and exemptions for small and micro enterprises with fewer than 50 employees and no more than EUR 10 million turnover.

22 Some of the authors have observed in their work and teaching that it is often tempting for such Small AI operators, and in particular providers, to operate with minimal, or even no domain expertise. Their team of developers may have formed around a small group of friends or associates, each expert in some technical discipline such as mathematics or computer science, or the hiring may not have been done with diversity of skills and knowledge in mind. And yet they may be working in an area in which their understanding is no better than that of a layman, with "big data" their only source of insight. In the case study of a young medical startup, Sappington and Serksnyte[48] show how an overly strong focus on expertise in engineering and mathematics can actively disregard many forms of domain knowledge, a scenario which some of the authors have directly observed too.

23 The lack of relevant expertise and resources in Small AI startups often means that they need to find alternative ways to build and maintain an ethical culture within their company.[49] However, this may incidentally lead to such companies believing they can try to contract out their responsibility, risk and liability for potential harm to the entities they provide their services or products to, and thus may take on zero moral or social responsibility themselves (and of course aim to avoid legal liability). Additionally, malicious and even non-malicious actors may occasionally operate with impunity by looking for legal and regulatory loopholes.[50] There are very few mandatory standards for them to follow, and many of the ethical frameworks[51] are merely voluntary. In their work with startups, some of the authors found that the notions of "malpractice", "negligence" or "duty of care" - let alone corporate social responsibility - appear not to have penetrated sufficiently into the team's (business) culture. Nor is the concept of "doing things properly" well-defined because there are no industry standards, accepted norms, and agreed-upon best practices to rely on.

47 cf. SIHCAI, *Artificial Intelligence Index Report 2023*, Stanford University (2023), https://aiindex.stanford.edu/ (last accessed Feb. 28, 2024).

48 Sappington, Serksnyte, *Who and What Drives Algorithm Development: Ethnographic Study of AI Start-up Organizational Formation, in* Proceedings of the Ethnographic Praxis in Industry Conference (2018), pages 245-263, https://doi.org/10.1111/1559-8918.2018.01206 (last accessed Feb. 28, 2024).

49 Sloane, *Ethics in German AI Start-Ups* (Dec. 22, 2022), https://dx.doi.org/10.2139/ssrn.4336465 (last accessed Feb. 28, 2024).

50 See Rainie et al., *Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm Within the Next Decade*, Pew Research Center (Jun. 16, 2021), https://www.pewresearch.org/internet/2021/06/16/experts-doubt-ethical-ai-design-will-be-broadly-adopted-as-the-norm-within-the-next-decade/ (last accessed Mar. 5, 2024), for some expert commentary on this and related problems.

51 For an overview see Ayling, Chapman, *Putting AI ethics to work: are the tools fit for purpose?*, 2 AI and Ethics 405 (2022), https://doi.org/10.1007/s43681-021-00084-x (last accessed Feb. 28, 2024).

Hence, attempts to punish small teams for not developing and maintaining their Small AI properly, or attempts to audit and check their output, would regularly fail. As a former member of the United States National Security Council, Roger Cressey, has argued: AI development "is the Wild West right now"[52] with nothing to keep these cowboys in check.

24 It may not seem as though Small AI should be of any broad societal concern, as their impact is rather restricted. However, in aggregate they do pose a risk to society through repeated and often undetected (hence not triggering a regulatory response) conduct, which might be described by a "death by 1000 cuts" metaphor. That is, numerous, perhaps primarily small-scale, impacts by Small AI on individuals, groups and even on how society functions that, however, in its combined effects can be severe. This might be in the form of people receiving inappropriate loan sizes or conditions, via the building of ineffective road networks, via the poor placement of hospitals, schools, and other essential services, via poor logistic networks, via the recommendation of inappropriate or inadequate products or services to people to buy or use, and so on. Aware that our categorisation involves simplifications, we suggest that AI systems that fit into the types described above share sufficient common denominators for the purpose of our risk and liability analysis to place them in one category. We call this "Small Team - Small AI".

2. Big Team - Big AI

25 Some large companies are implementing AI that regularly impacts a very large number of individuals, groups and even societies (usually beyond national borders), including large social media platforms. These companies will often be large-scale gatekeepers or decision makers, where in particular the cumulative effects of conduct by these entities are likely to have a profound effect on societies. This might be through providing a quasi-essential service such as a search engine, prolific social media platform, or online marketplace, all of which vast numbers of people have come to rely on and "could not do without". Or it could also be through the provision of platforms that other businesses build their own products and services around (such as web hosting, cloud computing, digital marketing, AI content generation, etc.). The impact of such conduct can be enormous: a Big AI used within a large organisation might change the way a significant portion of humanity acts, how people think, perceive the world, carry out work, or how they interact with each other. Even just one Big AI operator can, single-handedly, re-shape societies by impacting a significant proportion of the population. For example, a large search engine can change, or even distort, the information a significant proportion of the public can or cannot find. Or a video content recommender could recommend harmful or radicalising content to masses of people. Even if such impact is considered minor at the individual level, its sheer scale makes it significant.

26 The companies deploying Big AI are often extremely visible. Due to their large user base and ambitions, they cannot hide, and are not trying to. They aim to have as many users, customers or subscribers as possible, and their existence is usually well-known. They utilise AI and other algorithmic tools to filter and process vast amounts of data daily. This often leads to scenarios where Big AI influences how a lot of people think, and how they see the world.[53] As exemplified by the Cambridge Analytica scandal, this poses a political threat, as now one organisation can be the conduit for significant political influence[54] - without this being immediately apparent.

27 Serious problems might arise from the scale of their impact across society and the associated aggregated impact, rather than from individual actions or decisions carried out by such an organisation. For example, large platforms may use AI to recommend products or services to their users, control what they have decided to constitute hate speech, implement mechanisms to monitor other content the platform defines as 'harmful', or act to remove alleged copyright infringements. While blocking, de-listing or otherwise censoring access to any *one* article, post, video or other content may not have terribly harmful consequences for the author or society, large-scale conduct that deviates from applicable laws and societal norms on free speech is likely to have distortive effects on any democratic society, including by shaping - over time - what society considers to be acceptable as free speech. Even if no

52 Klar, *AI 'wild west' raises national security concerns*, The Hill (Mar. 8, 2023) https://thehill.com/policy/technology/3888433-ai-wild-west-raises-national-security-concerns/ (last accessed Feb. 28, 2024).

53 Susser et al., *Technology, autonomy, and manipulation*, 8(2) Internet Policy Review 1 (2019), https://doi.org/10.14763/2019.2.1410 (last accessed Feb. 28, 2024).

54 ur Rehman, *Facebook-Cambridge Analytica data harvesting: What you need to know*, 2019 Library Philosophy and Practice 2497; Berghel, *Malice domestic: The Cambridge analytica dystopia*, 51(5) Computer 84-89 (2018), https://ieeexplore.ieee.org/document/8364652 (last accessed Feb. 29, 2024); Wilson, *Cambridge analytica, Facebook, and Influence Operations: A case study and anticipatory ethical analysis, in* Proceedings of the 18th European conference on cyber warfare and security (Jul. 4-5, 2019), pages 587-595, https://www.proquest.com/openview/b2de0d9e0b088a0711696220cfa58d82 (last accessed Feb. 29, 2024).

one decision or action by Big AI may be particularly detrimental, at scale and in aggregate, its impact can shift societies in potentially unsustainable ways.[55] Once again, we see the harm here as "death by 1000 cuts". And even though the source here is centralised, the large-scale harm can arise from a compounding of minor harms.

28 In the name of enforcing ostensibly "good" rules such as the removal of harmful content, Big AI deployers might over-enforce such rules in the absence of any disincentives not to. Overall, this can lead to a systematic distortion of speech and communication, changing the entire information ecosystem. And while there are existing legal mechanisms to hold such organisations to account,[56] there are several challenges in doing so. As covered in Section A.II, there are difficulties in taking legal action against large entities deploying AI; even the Australian Federal Government needed to back down against Facebook.[57] There are difficulties in making the causal link between harm caused and the AI, let alone identifying AI was used in the first place. And there are difficulties in remedying the large-scale harms, given their type and sheer scale. All these suggest a new, tailored regulatory approach is needed. We call this "Big Team - Big AI".

3. Big Team - Small AI, and Small Team - Big AI

29 It might seem that specifying *team size* for AI is pointless, as in the examples we have given, Small Teams do Small AI, and Big Teams do Big AI. And as explained above, we assume that the size of the team generally correlates to the overall size and significance of any AI that team might produce or deploy. However, this is not always the case. Sometimes, Small AI is done by a Big Team, as might happen in a pharmaceutical research laboratory creating bespoke drugs for rare diseases. In such cases, a large organisation with substantial resources and many staff is developing an AI with a niche application which (directly) impacts only a relatively small number of people. Here it is probably safe to assume that the size and breadth of the team, combined with the very constrained application space, makes the use of AI in this setting rather low-risk for widespread societal damage, despite the costs for individuals potentially being extremely high. However, Big Teams producing B2B products and services for other organisations may have a significant, usually more indirect impact on society, not least because the consumers and other end-users of the respective products or services with inbuilt AI might not even be aware of the involvement of an AI system. This could, for example, occur in the pharmaceutical industry, either in the case of personalised medicine, or where AI is used as part of the pharmaceutical research process postulating which drugs to try (or indeed not try) to develop, and how.[58] We call this "Big Team - Small AI".

30 Conversely, a Small Team might be developing a Big AI, such as a startup that creates a widely-applied or widely-used product with significant uptake - such as, for example, an originally Small Team of OpenAI developers creating ChatGPT or the nowadays widely used Stable Diffusion from stability.ai. Here we might see a Small Team of developers create and run a product or service that directly impacts over 100 million people.[59] In this instance, we observe all the risks of Big AI described above, such as the potential for distortive effects on society that cannot be rectified ex post via means of dispute settlement. At the same time, we also see all the drawbacks of Small Teams, potentially including a lack of breadth, insight, and resources, thus making it hard for them to fully understand their impact space and hence to be able to address risks posed. Just as for Small Team - Small AI, in a Small Team - Big AI situation, one can also occasionally observe an abdication of responsibility by attempting to contract out potential harms through contracts and terms of service.[60] This combination of the risks of Big AI,

55 Arogyaswamy, *Big tech and societal sustainability: an ethical framework*, 35 AI & Society 829 (2020), https://doi.org/10.1007/s00146-020-00956-6 (last accessed Mar. 5, 2024).

56 See the discussion of different legal regimes in Section C; as well as Perc et al., *Social and juristic challenges of artificial intelligence*, 5 Palgrave Communications 61 (2019), https://doi.org/10.1057/s41599-019-0278-x (last accessed Feb. 29, 2024).

57 See Taylor, *Deliberate ploy: whistleblowers reveal why Facebook's Australia news ban included non-news sites*, The Guardian (May 28, 2022), https://www.theguardian.com/technology/2022/may/29/deliberate-ploy-whistleblowers-reveal-why-facebooks-australia-news-ban-included-non-news-sites (last accessed Feb. 29, 2024).

58 For an overview of the possibilities of AI usage in personalised (cancer) medicine we refer to Schork, *Artificial Intelligence and Personalized Medicine, in* Precision Medicine in Cancer Therapy 265-283 (von Hoff, Han, Jun. 2019), https://doi.org/10.1007/978-3-030-16391-4_11 (last accessed Feb. 29, 2024).

59 ChatGPT had more than 1 million users within 5 days, see Ul Haque et al., *"I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data*, arxiv (Dec. 12, 2022), https://doi.org/10.48550/arXiv.2212.05856 (last accessed Feb. 28, 2024), and well over 100 million users after only 2.5 months; for a rich discussion of its large-scale impact, see Leiter et al., *ChatGPT: A meta-analysis after 2.5 months*, arxiv (Feb. 20, 2023), https://doi.org/10.48550/arXiv.2302.13795 (last accessed Feb. 29, 2024).

60 One of the authors has extensive experience working

and the shortcomings of Small Teams, makes the use of AI in this setting susceptible to being extremely high-risk. We call this "Small Team - Big AI."

4. Risks and Progression in the Matrix of AI Types

31 We have established three key observations about the matrix of AI types:

- That "team size" and "AI size" are often, but not always, correlated,
- That "team size" and "AI size" can sometimes differ, and
- That the risk to society posed by the organisation increases as the AI size increases, and as the team size decreases. That is, in general terms, the riskiest AI potentially comes from "Small Team - Big AI", and the safest from "Big Team - Small AI".

32 What complicates matters further is that an AI operator can *move* to different squares in the matrix; their placement is not static. For example, a Small Team - Small AI operator might be using AI for niche research, and evolve or grow into a very large research team/organisation doing much more sophisticated and elaborate AI work, but broadly within the same niche area, thus becoming a Big Team - Small AI operator. Alternatively, a Small Team - Small AI might have a product or service that "takes off" and very rapidly gains 100 million users on a platform, or is deployed over a global supermarket chain with 100 million shoppers, thus becoming Small Team - Big AI. Or a Small Team - Big AI might scale up their operations and employ more developers and staff to cope with their increased impact and associated responsibilities, thus becoming a Big Team - Big AI. So, unlike regulating a use-case of AI (such as for insurance pricing) which is static with any given organisation, the AI type of an operator can change, sometimes rather quickly. Thus, one single regulatory approach is unlikely to be successful, and different strategies, corresponding to the dynamic developments described here, are needed. And as the context in which AI is deployed changes, there needs to be sufficient regulatory overlap to cover dynamically developing situations as they move between different squares in the matrix; whatever the development and deployment methodologies of the company are, and however they change, there should always be some, if not several, of these regulatory tools applicable. An AI operator should not be able to adjust its team size, and/or deployment size and space, in such a way that none of the regulatory tools (regarding Big/Small Team, and Big/Small AI) are in force.

with AI startups, providing advice to them for ethical and responsible AI development. They have dealt with over 25 AI startups, from 2020-2022. Part of their work involved investigating which entities within the startup, and which external entities (i.e., clients), would be responsible for the proper and safe functioning of the AI. In nearly all instances, across nearly all startups, this author was told some version of the statement "We contract out all responsibility for proper functioning, and subsequent liability for any malfunction, to the client".

B. Potential Solutions from a Practical Perspective

33 Having introduced the Big-Small axes (of team size, and AI size) to analyse the risk posed by AI systems, we now propose *two complementary regulatory tools which, when deployed together, become a solution* that we believe fits with our previous analysis. One is for Small AI: a form of strict(er) liability for the harm caused by AI. And one is for Big AI: a form of regulatory oversight by means of constant monitoring and cumulative penalties. This section outlines and addresses the technical, social and legal challenges for each solution; whereas the further issues of anchoring these solutions within a domestic legal and regulatory framework will be addressed in Section C.

I. What Can We Do to Address the Risks Posed by Small AI?

1. The Problem of Small AI

34 As explained, Small AI is often undetectable to wider society or impacted parties, its uses and domains are practically limitless and difficult to forward-predict, and it can be rapidly deployed in unregulated areas.[61] So regulatory frameworks assuming it is discoverable, or can be foreseen, or done on a per-sector basis, are unlikely to succeed. Small AI is often developed *exclusively* by technically-trained people (mathematicians, computer scientists, etc.) without domain expertise or insight into non-technical and ethical aspects of their work, assembled in Small Teams which follow a high-risk high-return business model and are prepared (or even expect) to fail. So regulatory frameworks relying on such organisations to self-regulate and develop AI responsibly, or that

61 One of the authors witnessed an AI startup team make the (paraphrased) comment "we only want to work in unregulated industries", as part of his work as an ethics consultant for Machine Intelligence Garage (part of Digital Catapult UK) in 2022.

they have a principal interest to survive, are unlikely to succeed.

35 Of course, there may already be ways to sue AI providers for damages resulting from the AI. However, as many Small AI organisations sell B2B, this "harm" manifests through the products and services offered by the B2C organisation that interacts with consumers or other end users, thus making it practically difficult to sue the AI provider, or indeed even identify that AI has been involved, its causal contribution to the harm materialised, and which entity might be responsible. The entity purchasing the AI tool could sue the Small AI provider for such damages, but in our experience most Small AI providers explicitly absolve themselves by contracting out all such risks and liability, thus (attempting to) place the responsibility for safe and effective operation of the AI, and any associated civil liability, on the AI deployer. Of course, whether such contracting out is enforceable is a different matter. While this type of contracting - even if passed on in agreements between the AI deployer and end-user - might not always serve to avoid tort liability vis-a-vis end-users (especially if mandatory consumer protection laws come into play), legal and practical difficulties remain. They include (1) difficulties for AI deployers or other actors at the end of a supply chain involving Small AI somewhere upstream to take effective redress action (if they have been successfully sued by a user for damages) against their contracting partners, or directly against the Small AI provider or developer upstream; (2) challenges in identifying (and proving) causal links, the materiality of a contribution, let alone breaches of a duty of care along complex supply chains; (3) as well as the willingness of the Small AI operator to go bust. And one must consider all the above in a cross-border context where actors along the supply chain potentially operate in different jurisdictions and subject their contractual relations to distinct applicable laws. This indicates that any sort of legal system or framework which relies on standard civil liability alone as a deterrent is unlikely to succeed.

36 In summary:

- One cannot detect the use of Small AI;
- One cannot forward-predict all uses of Small AI;
- One cannot regulate Small AI exclusively on a sectoral, industry-specific basis;
- One cannot assume Small AI entities have an inherent interest to self-regulate and act responsibly;
- One cannot rely on the survival desire of the legal form through which Small AI is deployed;
- One cannot assume that existing civil liability alone is a sufficient deterrent for Small AI.

37 In light of the above, one needs an alternative approach that convinces and incentivises those developing Small AI to address the concerns about AI risks. We outline a proposed solution below.

2. A Proposed Solution for Small AI

38 With the development of new, novel, innovative Small AI tools and applications, there is often only one group of people sufficiently well-placed to see and understand the operation, outcomes, and limitations of such tools: the development teams themselves. Thus, our proposal for how to address the problems created by Small AI will focus heavily on these development teams and the individuals within them.

39 Existing liability laws applied to AI, typically with cumulative requirements such as (1) needing to prove that harm was caused; (2) that there is a sufficient causal link to the operation of the AI; and (3) that those developing the AI acted intentionally or negligently, are simply too weak to be any serious behavioural deterrent to those developing Small AI. In cases of harm from Small AI, the existence of (individual) harm is often easier to determine and to observe when compared to Big AI, but proving a sufficient causal link between the harm and the AI's operation serves as a substantial burden. In addition, the typical requirement in conventional tort law to prove intent or negligence (in the form of the breach of a duty of care) on the part of the AI operator(s) then becomes a step too far, especially in an area such as AI where there are currently no well-established and agreed upon standards and norms.[62] So, established benchmarks for negligence hardly exist.

62 cf. Mittelstadt, *Principles alone cannot guarantee ethical AI*, 1 Nature Machine Intelligence 501 (2019), https://doi.

40 We therefore suggest the creation and adoption of a liability system that achieves the following:

- **Places a legal incentive on the Small AI operators to consider all realistically possible ways to prevent harm resulting from their AI.** This alleviates the need for regulators to be able to "review" all Small AI work - hence such an incentive, if effective, avoids the need for monitoring or auditing of the AI.
- **Extends such an incentive, and associated civil liability, to (some) individual persons involved.** This prevents such an operator from simply becoming a "phoenix" and dissolving itself to dissipate any liability.
- **Requires only proof of harm (existence, and causally linked to the AI), but not proof of negligent behaviour.** This removes the complexity for the legal system of understanding how the AI operates or was developed, and addresses the fact that standards and norms (and thus the concept of negligence) are not well-formed for AI development given how new it is. This pushes operators to do their own safety and quality assurance, to ensure their product is safe.
- **Allows the AI operator to reduce any regulatory penalties if it can prove that, despite the harm caused by its AI, it had taken all reasonable steps to prevent any foreseeable harm from materialising.** This draws from the aviation idea of proving it is safe to an acceptable level of risk (cf. Flannery 2001),[63] which is crucial to build faith and trust by the public in a new and uncertain sector. Given how unpredictable AI is, the ethos of "assume it is safe" needs to be broken.
- **Disallows the contracting out of the above defined responsibility.** This stops an AI provider from washing its hands of responsibility, and helps ensure product safety is always a concern for AI operators.
- **Applies to all forms of AI development, and is not limited by sector or industry.** This prevents Small AI operators from trying to hide in unregulated or underregulated sectors, and is necessary given that AI can be used in practically all industries now.[64]

41 By placing significant legal burden and liability onto AI operators, and in particular on AI providers who (historically) have seldom experienced such legal burden, this places the onus on the people who need to be paying the most attention to preventing harms from AI: the developers themselves. This will be an effective and meaningful change, and will assist in creating a cultural shift amongst those actually building Small AI.

42 These outcomes might be achieved through a variety of specific legal tools, some of which we present and discuss in Section C.I.

3. Additional Benefits of a Strict Liability Regime

43 Will a strict liability regime along the lines suggested above stifle innovation? No, we believe that this regulatory framework will reward the best - in the sense of socially most beneficial - innovators of Small AI. It will promote more rigorous testing of such AI, and encourage and commercially justify the diversion of resources to such activities in order to maintain or gain a competitive advantage. We believe this will lead to a safer, and thus better, product for customers, providers and investors alike. Customers are getting more for their money with a safer product, AI operators will find it ultimately more cost-effective and easier to maintain and update a product that has been thoroughly tested from the start, and investors will be at less financial risk when they invest in Small AI.

44 Objectors may argue that such an imposition delays production, reduces functionality, and increases costs, but this is not necessarily the case. Indeed, additional safety measures such as airbags in cars can be seen as an increase in functionality, giving manufacturers a competitive advantage. People are willing to pay more for safety, once they understand the consequences of not doing so. So those in industry who try to argue "No-one else is doing this, so why should we?" will soon find that their competitors *are* doing more to test and verify their product, thus saving on civil lawsuits, and gaining more customers with their safer product.

45 Such a liability system will also incentivise VCs and other investors to make safety, testing, and verification another of their key investment criteria. This turns formerly hidden costs into something public and measurable, thus encouraging companies to articulate, and investors to scrutinise, the

org/10.1038/s42256-019-0114-4 (last accessed Feb. 28, 2024).

63 cf. Flannery, *Safety culture and its measurement in aviation* (2001), https://api.semanticscholar.org/CorpusID:110462980 (last accessed Feb. 29, 2024).

64 Different sectors will naturally have different potentials and/or levels of harm. But the nature and scope of liability does not need to (and often cannot) be specified in advance as it will be highly contextual, while in principle the form and type of restitution or compensation should be relative to the harm caused.

planning, processes and budgeting that will go into additional safety checks. This also helps companies justify to investors, and investors demand of companies, specific resources and money being put towards the safety of the AI. Those companies with the weakest safety processes will pose the most risk for investors, and it is those companies that will be forced to change, or lose out on investment and thus be less likely to succeed. It will ultimately promote investments into companies that genuinely value AI safety and ethics, some of which are currently overlooked in the "arms race" of AI development.

46 With a strict liability regime as suggested in Section B.I.2 in place, this adds additional skills requirements in the corporate hiring process, and thus becomes a quality control mechanism for hiring. Each new hire will either need existing competences in AI safety, or require additional training on the job. It is usually much easier, and cheaper, to hire someone with an existing skill than to train them in it. Moreover, knowing someone has already learned that skill beforehand means it is probably a skill they are interested in and see value in, rather than something the company subsequently forces upon them, so they are likely to be more aligned with its relevance and purpose from the onset. And if this becomes a core part of the hiring process, it will encourage the education of AI developers early on, at university and/or at entry-level AI development jobs. This promotes innovation in education and training, as AI safety becomes a necessary skill for employment.

II. What Can We Do to Address the Risks Posed by Big AI?

1. The Problem of Big AI

47 The regulatory tools proposed above to combat Small AI cease to function *properly* when trying to deal with Big AI. Of course, having some tailored approach within liability law in place is certainly useful, and acts as *some* disincentive for bad behaviour and bad outcomes, so our proposed strict liability frameworks for Small AI operators should *continue to apply* for Big AI operators: just because their AI has a significant impact, does of course not mean they should be absolved of a stringent liability framework. However, as discussed earlier, several further issues arise. It may be difficult for an individual or group to directly challenge a Big AI operator in court, or to establish a causal link between a *specific* harmful event and the operation of Big AI. The harm inflicted by a Big AI might not be substantive enough in relation to a single individual or legal entity to bring a civil lawsuit, but at the same time the scale of the (systemic) damage inflicted on society might be too big to repair or compensate. Overall, the notion of post-harm responses, in particular in the form of liability for harm caused by Big AI would most likely be ineffective, or be pointless. Thus, more proactive and preventive additional measures need to be developed.

48 Unfortunately, forward-predicting how a Big AI operates and interacts with its surroundings is impossible, for several reasons. The training data is likely massive (and proprietary, or at least under the de facto control of few), and the ML training itself is often non-deterministic: repeating the same training on the same data is likely to produce different AIs.[65] While continuous (technical) audits can solve *some* problems,[66] they are fraught with difficulties owing to the nature of how AI is constructed (usually via a large matrix of "weights"). And the feedback loops Big AI may create, by changing bulk human behaviour and then "re-learning" on that, are often impossible to predict or foresee accurately.[67] Hence, any attempt to audit and understand how the AI will behave, pre-deployment, will be of little use. Thus, *real time monitoring* is needed.

49 It may be tempting to carry out such preventative and monitoring measures via "thorough checking" of all outputs of a Big AI. Of course, one cannot check all outputs and decisions by means of human review: a single Big AI might produce billions of outputs daily.[68] As described earlier, creating a separate digital entity alongside the Big AI to check all outputs simply expands the Big AI itself, and creating an independent digital checker would be practically impossible for most jurisdictions. Consequently, any attempt at comprehensive monitoring of the Big AI is probably futile. Therefore, some form of *selective monitoring* is needed instead.

50 At this point it may appear that keeping Big AI safe is impossible, as we can neither fully understand it,

65 Shumailov et al., *Manipulating SGD with Data Ordering Attacks*, 34 Advances in Neural Information Processing Systems 18021 (2021), https://doi.org/10.48550/arXiv.2104.09667 (last accessed Mar. 5, 2024).

66 Mökander, Floridi, *Ethics-Based Auditing to Develop Trustworthy AI*, 31(2) Minds & Machines 323 (2021), https://doi.org/10.1007/s11023-021-09557-8 (last accessed Feb. 28, 2024).

67 For a discussion of potential feedback loops in large-scale algorithms predicting violence see Hogan et al., *On the Ethics and Practicalities of Artificial Intelligence, Risk Assessment, and Race*, 49 Journal of the American Academy of Psychiatry and the Law 326 (2021), https://jaapl.org/content/49/3/326.long (last accessed Feb. 28, 2024).

68 In the case of OpenAI, GPT was receiving 2.5 billion prompts a day as of July 2025. See Silberling, *ChatGPT users send 2.5 billion prompts a day*, TechCrunch (Jul. 21, 2025), https://techcrunch.com/2025/07/21/chatgpt-users-send-2-5-billion-prompts-a-day/ (last accessed Jan. 5, 2026).

nor fully check it. But as discussed earlier, most Big AI harm comes from how it performs *in aggregate*[69] to pose systemic risks. This is particularly pressing in healthcare, where others have already called for a more encompassing approach to explainable AI.[70] Here we can make use of the product-service duality of AI, and view Big AI outputs as discrete items or *products, for the purpose of our regulatory control model suggested below* (that is, treat each output from the AI as a "product", just as one treats hammers manufactured in a factory as products). A Big AI might produce billions of outputs each day, and we wish to ensure that in aggregate these are acceptable. Thus, we can use established quality control mechanisms from manufacturing, where one periodically selects a small sample of objects produced, and then uses human checks on those to infer if the production line as a whole is working well.

51 In summary:

- One cannot expect substantive (compensable) harm to an individual from Big AI;
- One cannot always establish a causal link to personal harm from Big AI;
- One cannot expect that those harmed can afford to sue a Big AI operator;
- One cannot expect that full restitution is possible from suing a Big AI operator for any large-scale societal harm it causes;
- One cannot forward-predict all behaviour and impact of Big AI;
- One cannot conduct a full pre-deployment audit of Big AI;
- One cannot create another AI to test all outputs of a Big AI;
- One cannot test all outputs of a Big AI by hand.

52 In light of the above, we suggest that one needs an alternative approach. In our view, the best option is a constant safety monitoring system; one that is implementable, affordable, and detects problematic behaviour in almost real time. We outline our proposed solution below.

2. A Proposed Solution for Big AI

53 What we propose here is a constant monitoring system with human oversight for Big AI, coupled with a cumulative penalty system if significant problems are detected in its output. This is *in addition* to the liability system described in the previous section. We describe our proposal below. But first, we draw some simple analogies with existing systems in the world that most people are already familiar with.

54 Firstly, the concept of random testing of products in a factory production line is well-understood. For example, a factory filling water bottles has various ways of testing and manually inspecting some bottles, to infer whether the production line is working well. They might select a small fraction of the filled bottles each hour, to look for cracks, over/underfilling, etc. Then, if a certain percentage of these bottles are faulty, OR if the percentage of faulty bottles starts rising suddenly, OR if the percentage of faulty bottles fluctuates each hour throughout the day, it can be inferred that the production line has problems. This process can be formalised with existing statistical techniques in mathematics.

55 The overarching principle is to randomly sample objects, and ask "Is this object satisfactory?" for each one. It might be that the objects should be identical and conform to one single standard (e.g., filling the same bottle type). Or they might be checking several different products being produced together (e.g., different tiles made in the same factory). Or they could be checking completely different "objects" (e.g., a senior academic examiner, randomly sampling essays graded by more junior academics to check they are not over- or under-grading).

56 Second, the concept of a cumulative penalty system based on random sampling is also well-understood. For example, in many jurisdictions speeding cameras operate to catch drivers who exceed the speed limit. These cameras do not have *total* coverage of all drivers at all times, but take sufficient samples that a driver regularly speeding will eventually get caught, at which point, they are issued penalty points. The more often they get caught, the more points they accrue, and the greater the penalty administered (fines, driving bans, incarceration). Such a system aims to change the aggregate behaviour of drivers, by disincentivising speeding and administration of penalties,[71] and by removing drivers who repeatedly

69 For the aggregate effects of search engines see Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (Feb. 20, 2018), https://doi.org/10.2307/j.ctt1pwt9w5 (last accessed Feb. 29, 2024); for the potential aggregate impact of AI on the economy see Bughin et al., *Notes from the AI frontier: Modelling the impact of AI on the world economy*, 4 McKinsey Global Institute 1 (2018).

70 Ghassemi et al., *The false hope of current approaches to explainable artificial intelligence in healthcare*, 3(11) The Lancet Digital Health 745 (2021), https://doi.org/10.1016/S2589-7500(21)00208-9 (last accessed Feb. 28, 2024).

71 Corbett et al., *The deterrent effect of penalty points on speeding*

speed - rather than attempting to detect all instances of speeding. By using a monitoring system for Big AI, we can couple this with such a penalty system for the Big AI *deployers*.[72] The more often the Big AI gets "caught" doing something wrong, the worse the penalty.

a.) An Outline of the Constant Monitoring System with Human Oversight

57 The constant monitoring system has several components and steps, which we outline here. We start with some organisational prerequisites before moving on to the main monitoring loop involving seven steps.

Prerequisites

- The Big AI deployer that needs to be monitored must be informed that it is part of this monitoring, which will be carried out by what we call a 'Check Authority' (CA). The CA's competence and jurisdiction to monitor should have a clear legal foundation (e.g., in a tailor-made regulation) in order to ensure, and if necessary enforce, the cooperation of AI deployers in the monitoring process.

- The type of activity or outputs produced by the Big AI deployer need to be identified, as well as any relevant *existing* laws, standards or benchmarks that these outputs will be evaluated against. This may be, for example, testing if the AI is incorrectly flagging social media posts as hate speech, or scrutinising whether the AI is incorrectly taking down content as copyright-infringing. One will need to decide what is actually being tested for, and whether the concern here is false positives (such as AI wrongly identifying an upload on YouTube as copyright infringing), or false negatives (AI not detecting a copyright infringing upload on YouTube), or both.

- A team of individuals with relevant expertise will need to be assembled to undertake the testing.[73] These may be representing a regulatory body, civil servants of a government office, selected experts serving on boards reviewing the application of industry standards, or may even be judges, or other qualified people. The point is they are testing samples against *existing* laws relevant to the AI output in question (e.g., rules on hate speech, non-discrimination, competitive markets, or copyright), and already have the expertise to make evaluations or judgments on such matters.

Monitoring loop: sampling and testing decisions

1) The Big AI deployer will need to log all inputs, choices, decisions, and outputs it makes for a given time period to be determined. This could be something like "6 month rolling", to give the CA enough time to make all checks and have an appeals process, but not create a substantial data storage burden on the deployer. However, most AI deployers will already be storing their system's user inputs and AI outputs, hence the additional data burden should be negligible.

2) The Big AI deployer will then send through a (salted)[74] *cryptographic hash* (a small "unique fingerprint") of each decision or output to the CA. It preserves privacy rights by giving no information about the original file, but can be easily verified against the original file. This process also helps reduce the amount of data that needs to be transmitted to the CA. Often user inputs and the generated AI outputs will already be used by the companies to improve their product and to make new products. The consent potentially required for such data transfer could be part of the proposed legislation.

3) These hashes will need to be sent to the CA at regular intervals, in batches, by the Big AI deployer; potentially daily or weekly. To protect itself, the Big AI operator might take a "hash of each batch" and commit that to a large blockchain; this creates a very small file that can be stored on a blockchain in an efficient way. If the CA happens to lose some hashes in the future, the Big AI deployer is in a position to refer to the blockchain to verify what data had been sent.

4) The CA then randomly selects a small fraction of the hashes from the batch, and asks the Big AI deployer to send through the original files for checking. The CA cannot see in advance what data lies behind each hash, so cannot cherry-pick the "worst" decisions to test. Similarly, the Big AI deployer needs to hash all of their decisions/outputs, and cannot cherry-pick the "best" ones to send for testing. If it comes

drivers, in Proceedings of the Behavioural Research in Road Safety 2008: Eighteenth Seminar (2010), pages 175-195.

72 Our monitoring proposals are for Big AI *deployers* (not providers), as it is the deployment which determines the scale of impact. We place the onus on the deployers (who are "easy to see"), and they can then pass some responsibility on to others in the AI supply chain through contracting, careful choice of AI providers, setting of minimum performance standards, etc.

73 They act as a "human-in-the-loop". Implementing human-in-the-loop setups in practice come with their own well-studied challenges which we do not discuss here; for details see Chiodo et al., *Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility,* arxiv (Sep. 25, 2025), https://arxiv.org/abs/2505.10426 (last accessed Jan. 8, 2026).

74 A simple additional privacy step, to prevent any reverse-engineering of hashes back to outputs.

to light that a complaint is made against the Big AI deployer for an unfair or harmful decision, the deployer will need to present the output/decision, and corresponding hash, to the CA; if it was not one of the hashes originally sent over then substantial penalties are levied against the deployer.

5) The CA then presents this small fraction of AI outputs/decisions to the trained reviewers (judges, regulators, etc.) for individual scrutiny against existing laws.

6) The results of this human review process, which, in this system, are considered to be *a priori* correct (until proven otherwise), are then saved in a database. The "(time, correct/incorrect)" pair that is the result of the review process forms a time series for which the CA can then perform a change point detection using mathematical statistics.[75]

7) There are many ways to test if a time series has changed, but we outline three common changes that the analysis might look for:

- *Measure a change in the misclassification rate.* This simply measures the percentage of AI outputs/decisions determined as "wrong" by human checkers in a single batch. A threshold needs to be set initially, and if sufficiently many outputs are deemed "wrong", then the Big AI deployer is put on notice or given a penalty. One may introduce multiple thresholds with varying degrees of penalty.

- *Measure a change in the misclassification trend.* This measures how quickly the percentage of overruled outputs/decisions changes from batch to batch. For example, if most previous batches had roughly 1% errors, and then suddenly the errors start quickly climbing in successive batches to 3%, 4%, etc., then even though this might be under the tolerated error rate, the sudden uptick indicates a problem, and the Big AI deployer is put on notice or given a penalty.

- *Measure a change in the misclassification variance.* This measures how much the percentage of overruled outputs/decisions changes from batch to batch. For example, if most batches had between 0.9% and 1.1% errors, and then suddenly the errors in successive batches start fluctuating more wildly (say, 0.2%, then 1.8%, then 1.4%, then 0.3%), then even though this is all within the tolerated error rate, and the error rate overall is not increasing or decreasing much, this more extreme fluctuation may indicate a problem, and the deployer is put on notice or given a penalty.

58 The sampling and testing loop keeps repeating at regular intervals, and the costs for implementing this system should be borne by the AI deployers being monitored. Each time a "change" is detected, some sort of warning, penalty or restriction will be given to the deployer, which we detail below.

59 As part of this review mechanism, the Big AI deployer may appeal any of these human review decisions in a specifically designed dispute settlement mechanism. An updated result is recorded if such an appeal is successful, and the statistical analysis is (easily) recalculated. But given the cost of such an appeal and the fact that it would take a substantial number of human decisions to be overturned to change the outcome of the analysis, this is unlikely to occur often. In addition to this specific dispute settlement system, penalties awarded and related decisions that affect the Big AI deployer are subject to judicial review. We illustrate this whole process in Figure 3 below.

75 Many statistical tests and algorithms, including some machine learning based approaches, have been developed to detect changes in a time series. For an overview, we refer to Aminikhanghahi, Cook, *A survey of methods for time series change point detection,* 51 Knowledge and Information Systems 339 (2017), https://doi.org/10.1007/s10115-016-0987-z (last accessed Feb. 28, 2024).

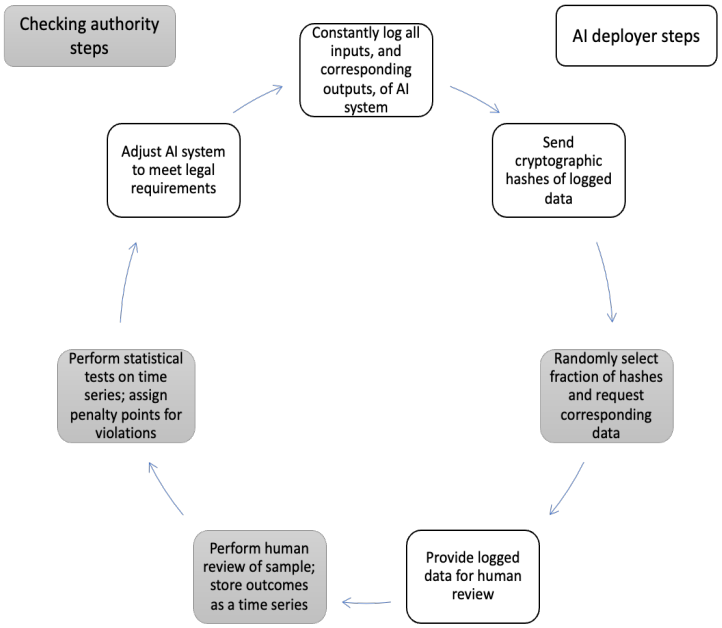

Figure 3: Intuition for constant monitoring with human oversight

Different types of Big AI to monitor

60 There are 3 types of AI that we might want to monitor as part of this Big AI approach, and the way we monitor each changes subtly.

1. Classifiers

61 These AI systems classify items into categories. For example, this could be an AI image recognition machine that scans bottles on a conveyor belt and classifies each as correctly/incorrectly filled,[76] removing the incorrectly filled bottles. So how would we monitor such an AI? From a business perspective, we want the number of bottles erroneously classified as incorrectly filled to be small. Reputationally (and legally), we want to minimise the number of bottles erroneously classified as correctly filled. Finding false positives vs false negatives can come with vastly different incentives and legal challenges. Here, we would independently sample bottles classified as incorrectly filled, and those classified as correctly filled, applying different thresholds for each group.

2. Predictors and recommenders

62 These AI systems make predictions (e.g., "It will reach 18 degrees on Tuesday") or recommendations (e.g., "These shoes will fit you") to help users make a decision.[77] They give more sophisticated outputs than classifiers, may not be falsifiable when the output is made, and are comparatively easy to manipulate.[78] It may require a long wait to observe if the output was correct. The errors produced here might be on a sliding scale, where we worry *how far* from "truth", or *how bad*, the output was. Our approach can still be applied, with a human reviewer identifying if the output was *sufficiently* good - based on either existing standards and legal principles, or those still to be developed. For example, the human might be a shoe fitter, giving their expert opinion against an AI shoe recommender.

3. Generators

63 These AI systems generate new content (such as text, image, video, etc.) based on an input prompt, like ChatGPT or AI generated news articles.[79] Reviewing such output is different from verifying "Is it correct?" as for classifiers, or "How wrong was this?" as for predictors. Understanding if the output was good or bad is often a human-subjective judgement, involving complex questions like "Is this *harmful* in some way?" requiring context-specific value judgements by the reviewer, with no immediate ground truth to check against, or process to falsify the output. These outputs require extensive human judgement, to evaluate the potential or materialised harm from the output, necessitating a human in the loop for any review process. However, as elaborated in Section B.II.4, our proposed monitoring system requires the human to evaluate outputs without significant reliance on contextual material or further external sources. Thus, it is less appropriate for determining harm which cannot be directly inferred from the output, or for evaluating broader societal harm (e.g., to a fair electoral process or democratic decision-making). Due to these inherent limits, lawmakers and regulators might consider adopting

76 Gizaw, Kebebaw, *Water Bottle Defect Detection System Using Convolutional Neural Network, in* Proceedings of the International Conference on Information and Communication Technology for Development for Africa (Nov. 28-30, 2022), pages 19-24.

77 Burke et al., *Recommender systems: An overview*, 32(3) AI Magazine 13 (2011), https://doi.org/10.1609/aimag.v32i3.2361 (last accessed Feb. 28, 2024).

78 Floridi et al., *How to Design AI for Social Good: Seven Essential Factors*, 26 Science and Engineering Ethics 1771 (2020), https://doi.org/10.1007/s11948-020-00213-5 (last accessed Feb. 28, 2024).

79 For a comprehensive overview see Cao et al., *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*, arxiv (Mar. 7, 2023), https://arxiv.org/abs/2303.04226 (last accessed Feb. 28, 2024).

a precautionary approach with regard to AI systems that pose risks of such type of harm, or rely on additional legal and regulatory tools, such as the risk assessments and audits relevant for large platforms under the EU's Digital Services Act (DSA).[80]

b.) An Outline of the Cumulative Penalty System

64 As well as monitoring, there needs to be some "consequence" that the Big AI deployer must face if the AI's performance is deemed unsatisfactory. One main benefit of a penalty system for Big AI, especially if deployed by a large corporate entity, is that many of the standard legal mechanisms and governance tools for controlling corporate behaviour, including fines, restriction of activity, and shutdown, usually *do* work. In general, Big AI deployers (usually comprised of a Big Team) will aim to comply, avoid (significant) fines, and are not prepared to be shut down, unlike Small Team operators who often are more immune to such threats, as discussed earlier.

65 The cumulative penalty system will trigger whenever a "change" is detected in the final (change-detection) step of the monitoring loop discussed above. A cumulative system's purpose is, within the principles and rules of the existing constitutional and other applicable legal orders (for example, the notions of proportionality, reasonableness and due process), to discourage any harmful or otherwise problematic behaviour of Big AI, without being so harsh as to dis-incentivise useful AI developments. Such a penalty system has two main components:

1) The size/severity of each change detected, with an associated number of "penalty points" issued to the Big AI deployer, and

2) The penalty issued to the Big AI deployer as it accumulates such "penalty points".

Penalty points

66 The exact changes and thresholds at which to issue penalty points need to be decided on, in advance. Mathematical insights can assist with making such decisions, but it will also require discourse on what constitutes bad outcomes and behaviour. This may change depending on the scale of the Big AI, the perceived risk to society (based on factors such as likelihood of occurrence and severity of harm) of the application domain and how quickly problems can be rectified (including whether the situation pre-harm can be restored, or whether financial compensation can be awarded). These are technical numbers but require extensive non-technical discussion and understanding. This is where the AI's "application domain axis" can naturally start to play a part, as the application domain often further determines many of an AI's inherent risks, and thus how severe the penalty point system should be (e.g., higher penalty points for riskier application domains).

67 Penalty points need to be proportional to the (potential) harm caused by the Big AI, pre-determined, formulaic, and the details made available to all parties beforehand. The number of points awarded, how long they stay on a deployer's record, and the consequences of accruing various numbers of points, need to be decided on. These decisions are made in tandem with the decisions about the changes to be monitored as well as the relevant thresholds. As indicated above, they have to be taken within the parameters of applicable constitutional and other relevant legal orders governing the conduct of providers of products or services that pose risks, or equivalent corporate behaviour.

Consequences of accruing points

68 The outcomes of incurring penalty points could be various different consequences or punishments, drawn from the range of legal tools available in a given constitutional and legal setting, to penalise corporations and individuals. The idea is that, as this is randomised testing, such consequences need to "ramp up", from very light consequences at the onset, to very severe ones when many penalty points are accrued. These could include:

- Warnings to the AI deployer.
- Public announcements about the AI's poor performance.
- Monetary fines (of increasing amounts).
- Restrictions/limiters on how much the AI is allowed to process (per day).
- Restrictions/limiters on how much trade or activity the deployer is permitted to do.
- Temporary suspension of the AI.
- Permanent suspension of the AI.

69 In order to "have some teeth", compliance with the consequences imposed should be monitored and if need be enforced. Failing to comply should be subject to responses that are equivalent to contempt of court.

70 At the same time, such penalty points should not be "lifelong", but rather should drop off or decay

80 See the discussion in Section C.IV below.

over (operation) time, so that a Big AI deployer is incentivised to, and rewarded for, improving performance. Overall, our suggestion is a proposal for measures that aim to prevent harm, rather than those that retroactively punish the occurrence of harm, administer recourse, or provide compensation. Thus, it should be recognised if the Big AI deployer has taken steps to reduce risk and harm, i.e., to enhance AI's safety. Any restrictions on the functioning of the AI system, or on the deployer itself, therefore should be lifted gradually.

Legal and social balance

71 Existing laws relevant in the domain the Big AI might operate in (e.g., large social media platforms and the requirement for the removal of hate speech) can be one-sided in terms of the aims pursued. They may focus on one specific objective, such as the prompt removal of hate speech, through public laws (such as the German Netzwerkdurchsetzungsgesetz, NetzDG) or even the criminal justice system. But they may leave impact on other interests and potential consequences (e.g., false positives and the accidental suppression of valid speech) as a matter for affected individuals to pursue in front of domestic courts. Thus, such Big AI deployers are inclined to unduly focus on one aspect (avoiding false negatives) to avoid potentially severe state sanctions, and be content to overdo it in the knowledge that there is little, if any, consequence for them in the other direction (creating false positives). For reasons discussed previously, we believe many Big AI deployers and platforms are not concerned enough by individual complaints brought against them, even if taken to domestic courts.

72 Our suggested approach allows a way to balance out how a Big AI deployer approaches problems. Having a state-enforced monitoring system collectivises the way society responds to the operation of the Big AI. At the same time, imposing the system's costs on the AI deployer ensures that risk prevention is appropriately internalised. Collective means of seeking redress from the AI deployer or platform via litigation (for example in the form of a class action lawsuit) is not a realistic option in many jurisdictions. And even where collective redress mechanisms are principally available under the domestic legal system, they are unlikely to succeed, or even occur, as the impacted and harmed parties would be hard-pressed to discover one another. A state-enforced monitoring system gets around this limitation and can be designed to incentivise (and if need be, pressure) the Big AI deployer to take account of all relevant interests and societal concerns that can be checked for in the monitoring process, and not just focus on complying with those legal requirements that appear most consequential, or are most likely to be invoked by the state or other actors it is most fearful of.

73 As well as there being stronger legal incentives on one side of the outputs that a Big AI deployer is producing (i.e., false negatives), there may also be enhanced social, regulatory or legal mechanisms that can address, or at least identify and highlight publicly, these same types of problematic output. If the Big AI operator consistently generates harmful output (for example, if an automated filtering system allows a lot of illegal content to be viewed by users on a social media platform), users will eventually complain and the media as well as existing regulatory authorities may pick it up, even if such occurrences are rare. This can lead to a similar adjustment of behaviour in the Big AI deployer that in turn then results in biased compliance. Again, few effective mechanisms exist to detect and/or address over-enforcement (e.g., in the form of the removal of otherwise legal content); there is unlikely to be media attention or public outcry if a legally-compliant video is removed from a platform. This is particularly true if the individual enforcement act is unlikely to be considered harmful in isolation (i.e., a platform removing a short video), even though such enforcement action may well, in aggregate, severely influence and distort the balance of interests within society, and might even shape societal behaviour (e.g., by normalising the absence of a specific type of content or expression that - while legal - tends to be caught in hate speech filters).

74 We suggest that our approach is particularly suited to identify and even prevent AI-induced imbalances in relation to different societal interests for which there is no existing mechanism to uphold the envisioned balance of interests - either because state enforcement mechanisms incentivise AI operators to prioritise a particular interest, or because the surrounding social and economic context serves to achieve the same result.

75 Of course, the sorts of errors the CA might try and check for will depend on *how* the Big AI is deployed; the fact that monitoring needs to happen is by virtue of the impact size of the AI, but *what* needs to be monitored for will be determined by where and how the AI is deployed. One might counter by asking why, then, are we proposing this additional burden on AI deployers *only* by their impact size? If a Small AI were being used in a "high-risk" situation such as in medical diagnostics or treatment, why would that only fall under our liability framework, whereas a seemingly low-risk Big AI such as a spam filter would (perhaps counterintuitively) attract an additional monitoring obligation? The reason here is that practically every Big AI will pose some *systemic* risk to society - in a way that Small AI is unlikely to do. Consider these examples: a medical Small AI will, by definition, only impact a (relatively) small number of

people, and while it may still cause significant harm to them, this will not pose a systemic risk to society. Conversely, a (malfunctioning) Big AI being used as a spam filter may let through billions of fraudulent or phishing emails, leading to recipients handing over or losing money to the spammers. If 1% of a billion recipients fall victim to this, on the scale of $1000 each, that would be a (relatively) small harm for each victim, but would raise $10bn which might go towards an organised crime syndicate, leading to their significant empowerment and additional harm to society through other sophisticated criminal activity given their new-found wealth. Here we see that the risk to individuals is relatively small, but the combined risk to society is enormous.[81]

3. Technical Advantages of This Approach

76 Our proposed regulatory framework using constant monitoring with a human-in-the-loop[82] provides several technical advantages over other existing suggestions (e.g., an AI checking other AIs). These include:

- Our suggestion is based on a mathematically mature and active research area, particularly when compared to the idea of AIs checking other AIs. Sampling and change point detection are both well-developed fields: early works of change point detection can already be found in the 1950s,[83] and many experts can help build such a system if they are given the legal framework and requirements. *[The approach builds on an "existing technique in industry".]*

- The approach is computationally efficient: in comparison to building and constantly retraining AIs to check AIs, it has comparatively low energy cost and requires minimal technology. *[The approach "does not require expensive technology."]*

- Because the review result used as the "ground truth" in the statistical analysis is figured out by human review, it is explainable and interpretable, and the mathematical tests and sampling are verifiable. *[The approach is "explainable, interpretable and verifiable."]*

- Human in the loop: the human can spot patterns and raise additional alarms. Tacit and expert knowledge is not neglected (as would be the case for a purely technical solution) but actively used and encouraged. *[The approach "retains a human in the loop."]*

- The technical system requires little maintenance and (almost) no "retraining". As long as the threshold parameters do not change, we can simply let the system run. When a company appeals decisions, the test statistic of the statistical analysis needs to be recomputed, but this task is simple and could be automated. *[The approach "requires minimal maintenance."]*

- Changes can be detected purely based on the output, and independent of the AI: it does not matter if the operator updates its AI, and the system does not need to know anything about the internal operation of the AI. *[The approach is "technologically neutral."]*

- The system is relatively easy to adapt to new laws and regulations. For example, when the definition of hate speech changes, the new law only affects the human reviewers who need to be retrained for the new laws and regulations; the mathematical analysis does not need to change. This is not true for "AIs checking other AIs": such AIs will need to undergo much more expensive retraining when the definition of hate speech changes. Our solution only needs to re-adjust the human review mechanism. *[The mathematics is "law neutral."]*

- The approach is easy to adapt to different countries, cultures, values and norms, because different jurisdictions or areas that have their own laws can simply choose their own parameters and thresholds for what is considered unacceptable. *[The approach is, as a matter of principle, "culturally and jurisdictionally neutral."]*

- The system can be adapted for different risk classifications and different problems. While the relevant thresholds, parameters, etc., need to be changed, the overall architecture stays the same. *[The approach is "risk neutral."]*

- The parameters required for the statistical analysis and the human analysis can be chosen without extreme difficulty. It is not easy, but it is not nearly as impossible as "building an AI that can check another AI". Any such AI would also require us to answer many more technical, legal and moral questions. *[The approach is "not hampered by extreme difficulties."]*

- The mathematics and the human checks are decoupled: this means that for different problems different mathematical tests can be used, so the

81 One can only imagine what harm an organised crime syndicate could inflict with $10bn. The example of prohibition in the US in the 1920s brought rapid and vast wealth to organised crime syndicates, enabling them to flourish and establish themselves for decades to come.

82 For details on what a human-in-the-loop setup is, see Chiodo et al., *Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility,* arxiv (Sep. 25, 2025), https://arxiv.org/abs/2505.10426 (last accessed Jan. 8, 2026).

83 e.g., Page, *Continuous inspection schemes*, 41(1-2) Biometrika 100 (1954), https://doi.org/10.1093/biomet/41.1-2.100 (last accessed Feb. 28, 2024); *Control charts with warning lines*, 42(1-2) Biometrika 243 (1955), https://doi.org/10.1093/biomet/42.1-2.243 (last accessed Feb. 28, 2024).

approach is easy to adapt in the future. *[The approach is "future proof and adaptable."]*

- The mathematics can be implemented: this is not a multi-billion parameter AI that only a few people can implement. This is feasible to build, even in a jurisdiction with limited resources to spend. *[The approach is "feasible to implement."]*

- There are not too many Big AI deployers (compared to Small AI operators), which reduces the effort needed. And the monitoring can scale up/down smoothly depending on the size of the AI; more frequent monitoring (daily batches) for very large, and less frequent for smaller (monthly). *[The approach "scales easily with the number and size of such AI systems."]*

4. Limitations of This Approach

77 The "incorrect", or "problematic", decisions need to occur sufficiently often for this approach to be effective. If only a very small fraction of decisions being tested is incorrect, then a sampling-based approach would require a large sample, in order for the approach to work well. The exact definition of "very small" will depend on the application domain, the enforcement capacity of the Check Authority (CA) and other factors.

78 As already hinted at above, our constant monitoring approach relies on the human reviewer being able to evaluate the outputs and decisions of the AI shortly after they happen. The longer the time lag between the AI generating its output and the consequences of that output manifesting, the greater the time lag in our approach, thus making the approach less effective as a preventative measure. A scenario where there is a significant delay between the AI producing an output and it being determinable that that output was undesirable, leaves the AI with a long window to keep generating undesirable outputs before the review mechanism is able to detect and respond. That is, the AI will keep running and keep producing bad outputs until outputs are realised to be bad. In summary: the system works best if the AI decision can be reviewed and is falsifiable *quickly,* and the false positives and/or false negatives are sufficiently dense in the set of decisions.

79 That said, there will be scenarios where Big AI deployers choose to use their AI in settings where the output is unfalsifiable. Examples could include uses in the criminal justice system, where AI systems might be used to determine the chance of a convicted criminal reoffending, and thus factor into a sentencing process, as happened in the US with the system known as COMPAS.[84] Given that the "false positive" side of this decision process is completely unfalsifiable (an incarcerated criminal lacks the opportunity to "prove" they would not have reoffended if released sooner), our constant monitoring approach would not be applicable for such decisions, as there is no "absolute truth" to refer to. In *very limited* circumstances it may be possible to adjust our system, and have the checker see if the AI followed or violated certain procedural standards. However, since advanced AI systems often reason differently from humans, such examples may be limited to specific interpretable and explainable algorithmic decision systems, such as those using decision trees. But in general, it is inadvisable to make use of AI in unfalsifiable scenarios such as the one described above, as one of the (many) criteria for responsible and safe deployment includes falsifiability.[85]

80 However, humans frequently make unfalsifiable decisions. If an AI were to be used to replace a human making unfalsifiable decisions (the authors do not advocate for this), then nonetheless our constant monitoring system could be useful. Rather than checking against some sort of "absolute truth", the checker might make their own expert determination of the decision, to compare to the AI output. Thus, our system can be used to ascertain if the AI is, *on average*, performing as well as a human in the same situation.[86] This can also be refined, for example, in our criminal sentencing example, it might check how effectively an AI is performing on certain demographics of particular concern.

81 Our proposed constant monitoring system cannot answer all possible questions about the (mal) function of the Big AI. It is very effective, and quick at detecting *if* there are problems with the AI but does not necessarily explain *why* such problems have occurred. For example, it cannot answer questions such as:

- Why was there a change in the AI output?

- Why has the system broken down?

- Who is affected by the system?

82 However, we assert that it is the responsibility of the Big AI deployer to figure out and rectify such problems, just as it is the responsibility of car drivers

84 Rudin et al., *The Age of Secrecy and Unfairness in Recidivism Prediction*, 2(1) Harvard Data Science Review (2020), https://doi.org/10.1162/99608f92.6ed64b30 (last accessed Mar. 27, 2024).

85 Floridi et al., *How to Design AI for Social Good: Seven Essential Factors*, 26 Science and Engineering Ethics 1771 (2020), https://doi.org/10.1007/s11948-020-00213-5 (last accessed Feb. 28, 2024).

86 Noting that better average-case performance does not imply better performance in each instance.

to understand what they are doing wrong in their driving and change their behaviour accordingly. If the Big AI deployer is incapable of identifying the source of such problems and/or is unable to rectify them, or even goes so far as to claim that such actions are not possible, then this brings into question whether this AI should be permitted to continue to exist and operate. After all, it is the responsibility of those deploying such high-impact AI to ensure their tool is safe, rather than it being left to the rest of society to fix the problems (or pick up the pieces when harm is done). The overriding principle here is simple: If you cannot use your technological solution *safely,* you should not use it at all.

83 Lastly, the monitoring undertaken by the CA most likely cannot appropriately test for more diffuse risks for broader societal interests such as a functioning democracy, a system free from bias or from systemic discrimination of specific groups. As already indicated above, this is because the legal principles and rules which operate to uphold such broader interests are complex and usually will require significant and careful balancing of various interests, detailed and fact-specific determinations, and usually involve quite some discretion on the side of the (usually judicial) body in applying the relevant rules and principles to the concrete facts at hand. While all of this of course forms part of what trained lawyers (especially judges) do, this process requires time to hear arguments from all sides, proper representation of the interests concerned, experience drawn from similar cases, expert evidence on the facts and careful deliberation on the side of the decision-maker. It would be illusory to expect this to be replicable at scale in the context of constant monitoring, even if undertaken by trained experts on the point of law at issue. While we refrain from drawing an exact hard line on the types of legal issues that can be reviewed by a CA, the principle set out above stands: the more an AI output poses a concrete systemic risk of harm to clearly understood protected interests such as (human, animal or plant) life or health, to property, or other interests legally recognised in concrete and defined terms, the easier this can be made subject to our proposed monitoring system. On the other hand, the vaguer or more diffuse the legal interest at stake is and the more remote or indirect the risk posed by the AI system, the less helpful checks along the lines we propose will be. Alternative mechanisms, such as the comprehensive risk assessments by and external audits for very large online platforms and search engines under the EU's Digital Services Act (DSA),[87] may be considered to address more remote or indirect risks and/or potential harm to more diffuse and vaguely defined legal interests.

5. Additional Benefits of Constant Monitoring with Penalty Points

84 Is constant monitoring with penalty points a reasonable imposition for Big AI deployers? We believe that our approach provides more flexibility than existing regulations (e.g., the GDPR, or the EU AI Act) to control AI and data. Those regulations do not permit failure, in the sense that a single violation can lead to significant penalties for an AI operator. By comparison, constant monitoring allows, and even expects, some degree of failure, within certain tolerated levels. And even after a "failure", the penalty point system is cumulative, and early on might only inflict something as minor as a warning to the company. This is much less arduous on companies than regulations that require 100% adherence to stated rules with significant penalties for even the smallest violations, thus giving the organisation and its developers room to breathe and learn.

85 Our proposed system with constant monitoring and associated penalties is also much more likely to be accepted by those deep within the AI development and deployment process. Unlike the very rigid emissions testing regime for nitrogen oxides (NOx) which was routinely cheated by car manufacturing companies around the world, our proposal breaks down the barrier of acceptance of failures and enables developers to discuss failure openly. The regulatory constraints are no longer clear "lines in the sand" which cannot be crossed, but rather performance objectives to try and stay within as often as possible. The penalty points system actually gamifies the regulations, and helps developers see these as their own scoring system for how well they have done, turning failing into an opportunity to improve. This can help instill a workplace culture of seeing the regulations as an aid and working with them, rather than as a solid barrier and trying to work against or around them.

86 Existing regulations are often outcomes-based punishment systems, with very little feedback to developers. Instead, our monitoring + penalty points system can be viewed as an institutional training regime, supporting trial and error iteration for product improvement. This promotes innovation by creating a mechanism for constant improvement. It assists with the survival of the company by detecting problems that may lead to disastrous outcomes, before they manifest. And it fosters improvement and socially beneficial innovation by giving even ethically untrained developers and managers a starting point to build from and a long-term

87 For a discussion on how the legal tools of the DSA might serve as a useful framework for implementing the monitoring system considered here, see Section C.IV.

objective to aim for. Indeed, this system provides developers with incremental, external feedback on their product's performance, which is something they might otherwise have no access to. It helps them, and their product, improve, through a form of reinforcement learning for their own institution and its people. And if explained to developers in this way ("reinforcement machine learning on machine learning developers"), they will almost certainly respond well to it as a useful method for improvement.

87 Our suggestions here also stand up well to regulatory evaluation criteria set by others as well. What we have provided is an innovation-*friendly* approach to regulation, as discussed in the 2020 report of the UK Department for Business, Energy & Industrial Strategy.[88] Our approach of monitoring and penalty points matches well with 4 of the 5 key criteria set out in that report (cf. pp. 5-6). Namely, our approach:

- Provides regulatory advice to innovators,
- Supports experimentation and testing of innovations,
- Streamlines regulatory approvals for innovators, and
- Sets regulatory challenges to drive innovation.

88 Our proposed regulatory framework is as business-friendly as anyone could possibly hope for, and simultaneously one of the best possible ways to maintain the large-scale safety of Big AI. It can keep society safe, and keep businesses engaged.

C. Potential Solutions from a Legal Perspective

89 In this Section, we elaborate how the systems described in Section B can be anchored in existing and upcoming legal frameworks, resulting in deployment that aligns with existing legal systems and tools to the maximum extent possible. For this purpose, it is examined whether and how the strict liability regime suggested above for both Small and Big AI could be introduced, and whether the constant monitoring and penalty system for Big AI is reflected in existing liability law, market regulation and regulatory law. Within the analyses of these areas of law, random testing and the penalty points system are considered separately from one another, since from a legal perspective they constitute obligations of different types and intensities.

I. Amending Liability Rules (for Small and Big AI)

1. Introduction to Liability Law

90 Liability law operates as an indirect approach to regulation.[89] In contrast to direct ex ante regulation, imposing a set of obligations which must be met before an AI system may be put on the market (see Section C.V), there are no immediate instructions for action and hence no direct command to implement a monitoring system. The regulatee can, however, be held liable for damage that occurs. Assuming that the regulatee as a rational actor has an interest not to offer reparation or pay damages and hence aims to avoid liability, by establishing liability rules, the regulator influences the regulatee's behaviour indirectly by creating an incentive to act lawfully in order to avoid liability. In this context, a system of constant monitoring could hence be adopted as a prerequisite for compliance with the duty of care - so that monitoring becomes a necessary element to avoid liability. Furthermore, looting strategies with small entities prone to bankruptcy in case of liability can be avoided by combining liability rules with compulsory liability insurance (third party insurance).

91 It should be emphasised that our contribution in this article is *not* to offer new solutions or proposals for tort liability within Law (apart from suggesting strict liability for certain AI systems), but rather to situate our suggestions for Small and Big AI within the existing legal framework and examine whether this framework can accommodate our suggestions set out in Section B.I. For this analysis, it is essential to differentiate between different liability regimes: fault-based liability, presumed fault and strict liability which are briefly introduced below.

a.) Fault-Based Liability

92 Fault-based liability means damage is compensable only if the injured party proves it was caused by the injurer's wrongful behaviour, such as failing to observe a duty of care. Random testing, as part of the

88 UK Department for Business, Energy & Industrial Strategy, *Regulator Approaches to Facilitate, Support and Enable Innovation. Final Report. BEIS Research Paper Series, Number 2020/003* (Jan. 2020), https://assets.publishing.service.gov.uk/media/5e2efd76ed915d1f26a7427f/regulator-approaches-facilitate-support-enable-innovation.pdf (last accessed Feb. 28, 2024).

89 Zech, *Liability for AI: public policy considerations*, 22 ERA Forum 147 (2021), https://doi.org/10.1007/s12027-020-00648-0 (last accessed Feb. 28, 2024). In California, the new AI Bill explicitly mentions civil penalties and monetary damages.

monitoring system in Section B.II.2, could form part of this duty. If an operator neither monitors its AI nor performs random tests, it breaches this duty—creating an indirect obligation to test.

93 Applying fault-based liability to AI is challenging. The opacity and complexity of AI systems and their supply chains make it hard to prove defectiveness, causality, or wrongful conduct. For instance, a user receiving unsuitable insurance recommendations may not know whether AI was involved or where the error originated. Legislators or courts could address this via rebuttable presumptions of fault (see Section C.I.1.b).

94 AI's learning capacity further complicates attribution. An autonomous car might cause damage through behaviour learned independently, making it difficult to link fault to the provider. Proving fault under current law is thus often impractical. Small AI adds hurdles of invisibility and inaccessibility: startups may supply AI embedded in other products, leaving users unaware of the actual operator. While similar issues exist outside AI, the stakes are higher here (see Section B.I.1).

95 Big AI introduces power imbalances. Injured parties rarely have resources for lengthy, expert-heavy litigation. Class actions can help by pooling claims, but they falter when harm is systemic, minor, or non-pecuniary, and jurisdictional issues often arise. While class actions may incentivise Big AI operators to internalise harm—especially when paired with presumed fault or strict liability—they cannot resolve all challenges. Preventive measures like constant monitoring remain essential.

b.) Presumed Fault

96 In order to solve the aforementioned problems, tort law at times turns the burden of proof so that in specific scenarios, the injurer's (in this case the operator's) fault is presumed. Even in cases where fault is presumed, damages are only compensable if it is proven that these damages were caused by the injurer's reproachable or wrongful behaviour. It is, however, not presumed that the injurer's wrongful behaviour has led to the damage. Effectively, with this mechanism the burden of proof is shifted from the injured party to the alleged injuring party which needs to show that it has complied with relevant obligations, such as a duty of care. In order to escape liability, the alleged injurer therefore can make use of an opportunity for exculpation, by offering evidence that it has complied with the relevant duty of care (i.e., by showing that as AI operator, it did perform the required random testing).

97 With regard to presumed fault, however, three central issues remain unsolved: First, the above-described problems associated with power imbalance, invisibility and inaccessibility are not tackled. Second, the burden of proof only shifts for one aspect; all other aspects (such as that the damage has been caused by AI) must still be proven by the injured party. However, an additional presumption of causality might be introduced (like the new liability Directive envisages for complex products)[90] - which in turn would lead to a significant further reduction of the requirements for liability. While AI operators might of course argue such extensive presumptions of liability would create significant disincentives for the development of new technologies, we would counter that extended liability presumptions can actually incentivise more socially-beneficial technological innovations, and prevent a 'race to the bottom' in terms of safety and security.[91] Third and relatedly, the injured party will usually not have access to the relevant information and might not have sufficient expertise to be able to assess how and what caused the damage.

c.) Strict Liability

98 Strict liability makes damage compensable if caused by the operator's AI, regardless of wrongful behaviour or due care. It is an exception to fault-based liability, applied in hazardous contexts or where risks commonly materialise (e.g., car operation or dog ownership). Its justification lies in the risks of providing AI systems. Key advantages include internalising external risks and enhancing legal certainty. Since liability applies even when due care is observed, operators must weigh benefits against risks before market entry. Legal certainty improves because proving wrongful behaviour becomes unnecessary.

99 Strict liability signals that AI operators are accountable for AI-caused harm. It does not necessarily deter AI development—goods under strict liability remain widely used; people still drive cars and own dogs. As long as damages are insurable, strict liability is no disincentive. For businesses, legal certainty is crucial for profitability. Because due diligence is irrelevant here, random testing becomes voluntary. Operators will adopt monitoring or testing mechanisms they find cost-effective to reduce errors. Insurance providers may require such mechanisms as conditions for coverage, creating incentives for safety optimisation. If these mechanisms effectively reduce harm, they serve both operators and insurers.

90 See art. 10 sec. 4 Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, hereinafter: Product Liability Directive.

91 See the discussion in Section B.I.3 and B.II.5.

d.) Recent Legislative Activities

100 In September 2022, the European Commission (EU COM) published a Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive).[92] However, on 11 February 2025, the AI Liability Directive was withdrawn. The AI Liability Directive was intended to refer extensively to and complement the AI Act (see Section C.V.3). It is not yet foreseeable whether there is going to be a specific European liability regime for AI, and what it will look like in the future.

101 The 2022 Proposal shows that the EU COM did not envisage strict liability for AI, but intended to design a complex system of disclosure obligations and rebuttable presumptions of non-compliance and causation. The Draft Directive sought to lay down rules on the disclosure of relevant evidence by those that have it at their disposal, for the purposes of establishing liability. This measure was intended to provide an additional incentive to comply with the relevant requirements laid down in the AI Act to document or record the relevant information. In addition to that, potential claimants would have a right to request a court to order the disclosure of relevant evidence before submitting a claim for damages, if the potential claimant presents facts and information sufficient to support the plausibility of a claim for damages and it has made a prior disclosure request to the defendant. In case the defendant does not disclose the relevant information it is rebuttably presumed that the respective obligations under the AI Act have been breached. In addition to that, the causal link between the fault of the defendant and the output produced by the AI system is rebuttably presumed. The causal link between the AI's output and the damage must, however, still be proven by the claimant.

102 The proposed AI Liability Directive was to follow a minimum harmonisation approach which allows claimants in cases of damage caused by AI systems to invoke more favourable rules of national law. EU member states could, therefore, maintain reversals of the burden of proof under national fault-based regimes, or strict liability regimes. However, since the proposal has been withdrawn it is futile to speculate at this point as to what liability rules are likely to be developed by the EU and other actors, and how they will apply to AI in the future.[93] If anything, the discontinuation of the EU's draft AI Liability Directive suggests that the current pro-AI innovation climate makes the introduction of express presumptions (which lead to stricter liability standards for AI) politically not feasible.

2. Implementation of a Strict Liability Regime

103 As outlined in Section B, for Small AI the introduction of a liability relying solely on causation might be a necessary and sufficient way of addressing the associated risks. Such an approach was taken by the European Parliament's proposal[94] for the introduction of a new liability regime for AI.[95] However, the European Commission focused on an extension of the existing product liability (which relies on defectiveness) to AI, instead of introducing a strict liability rule.[96] Here, only some of the issues associated with the introduction of a strict liability regime shall be presented.

a.) Definition of Specific Risk Sources

104 The first problem is to properly define the relevant AI systems, or in general the relevant risk-causing activities, that trigger the strict liability. The EU proposal relied on high-risk AI systems to trigger strict liability (other AI systems being subject to fault-based liability with a presumption of fault). Similarly to the AI Act, the question of which

92 European Commission, *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*, COM(2022) 496 final (Sep. 28, 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496 (last accessed Nov. 12, 2025).

93 In 2022, the UK's Financial Conduct Authority (FCA) released a new Consumer Duty (PS 22/9) which sets out rules for financial institutions. Amongst other things, the Duty obliges financial institutions to mitigate harms and allows authorities to check whether these obligations are met and, in case they are not, to impose fines. With their new Consumer Duty, the FCA imposes a strict liability regime for severe harms as well as a duty to take mitigating measures to avoid liability. For more information see UK Financial Conduct Authority, *PS22/9: A new Consumer Duty* (Jul. 27, 2022), https://www.fca.org.uk/publications/policy-statements/ps22-9-new-consumer-duty (last accessed Feb. 28, 2024).

94 Annex B of the European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence, 2020/2014(INL), 2021 O.J. (C 404) 107.

95 cf. Zech, *The European Parliament proposal for a regulation on liability for the operation of artificial intelligence systems*, 14 Zeitschrift für Geistiges Eigentum/Intellectual Property Journal 293 (2022), https://doi.org/10.1628/zge-2022-0022 (last accessed Feb. 27, 2024).

96 See below Section C.III; cf. Wagner, *Liability Rules for the Digital Age – Aiming for the Brussels Effect* –, Working Paper No. 10, Forschungsinstitut für Recht und digitale Transformation (2023), https://www.rewi.hu-berlin.de/de/lf/oe/rdt/pub/wagner-liability-rules-for-the-digital-age-1.pdf (last accessed Feb. 28, 2024).

AI systems fall under the high-risk category is answered by referring to an Annex which can be amended by delegated acts (i.e., by the European Commission being the executive body without further involvement of the legislative body of the European Union).

b.) Determination of the Responsible Parties

105 Another issue is to determine the parties responsible for damages caused by the AI (aside from proving the causal link between the use of an AI system and the damage). This can be either the provider (programmer, trainer) or the deployer of such a system. The Parliament's proposal focuses on the deployers.[97] However, an argument can be made that with increased automation and autonomy the deployer's influence on the behaviour of a system diminishes. Therefore, rather the provider of such a system should be addressed by a strict liability rule (liability following risk control).

c.) Compensable Damages

106 The third issue to address is the scope of the damages that may be compensated. Like with most strict liability rules, in exchange for a mere requirement of causation, the scope of liability in the Parliament's proposal is limited to personal injury or damage to property. This means that sole (or pure) economic losses are not encompassed. For example, a claimant might be compensated for damage to their health caused by AI, but not for missing out on profits otherwise earned, or for not being considered for a job. Having said this, compensation for personal injuries explicitly includes substantial non-material damages (cf. art. 6 sec. 1 lit. a Product Liability Directive: "medically recognised damage to psychological health"). This still excludes injury to a person's honour (as in the case of defamation caused by AI), but definitely includes psychological injuries suffered through AI (e.g., caused by a chatbot's responses).

3. Product Liability for AI

107 The EU is set to go another route: extending the existing product liability to software and AI. This is a central part of the recent amendment to the existing Product Liability Directive which focuses on the producer's responsibility.

a.) AI as a Product

108 Whereas under former product liability law only movables (and electricity) can be subsumed under the definition of a product, this changes under the new rules. According to art. 4 sec.1 of the new Product Liability Directive "'product' means all movables, even if integrated into another movable or into an immovable. 'Product' includes electricity, digital manufacturing files, raw materials and software". Although it may be argued that AI systems do not constitute software, recital 13 of the Product Liability Directive clearly indicates that at least the *applications* of AI systems are intended to be included ("[...] Software, such as operating systems, firmware, computer programs, applications or AI systems, [...]").

b.) Defectiveness of AI

109 Product liability relies on the defectiveness of a product (in addition to the causation of a damage) as the trigger for liability. Defectiveness might be seen as a lack of diligence on behalf of the producer in disguise.[98] According to the new Product Liability Directive, a product shall be considered defective when it does not provide the safety which the public at large is entitled to expect. This has to be proven by the claimant. However, the burden of proof is eased because the claimant only needs to prove the objective defectiveness instead of the underlying behaviour of the producer. When it comes to AI, proof of defectiveness might be problematic due to the lack of transparency inherent in many AI systems (see Section B.II.1). The new Directive addresses this problem by introducing a presumption of defectiveness under certain circumstances (as set out in art.10 sec. 2). In addition, the causal link between defectiveness and damage should also be presumed where the damage caused is typically consistent with the defect (art. 10 sec. 3). Finally, in cases where proving the defectiveness or the causal link presents excessive difficulties due to technical or scientific complexity, a presumption under art. 10 sec. 4 comes into play.

97 For the precise determination of the responsible parties with the concept of frontend operator and backend operator see Zech, *The European Parliament proposal for a regulation on liability for the operation of artificial intelligence systems,* 14 Zeitschrift für Geistiges Eigentum/Intellectual Property Journal 293, 298 (2022), https://doi.org/10.1628/zge-2022-0022 (last accessed Feb. 27, 2024).

98 Wagner, *Robot Liability, in* Liability for Artificial Intelligence and the Internet of Things 27-62 (Lohsse et al., 2019), https://doi.org/10.5771/9783845294797 (last accessed Feb. 28, 2024); Zech, *Liability for AI: public policy considerations*, 22 ERA Forum 147 (2021), https://doi.org/10.1007/s12027-020-00648-0 (last accessed Feb. 28, 2024).

c.) Compensable Damages

110 Similar to strict liability regimes, compensable damages are restricted to material losses (cf. art. 6 sec. 2 of the new Product Liability Directive), meaning death or personal injury (including medically recognised harm to psychological health), harm to property and, as a new addition, loss or corruption of data. Pure economic losses, resulting for instance from discrimination against the injured party, are not covered. This is one of the important remaining functions of national tort law where under certain circumstances pure economic losses might be covered as well.

d.) Product Liability and the AI Supply Chain

111 Product liability addresses the producer of AI systems. This regulatory approach therefore shifts the focus of legal responsibility from the deployer (user) level to the provider level. With regard to the AI supply chain, this constitutes an important shift of responsibility upstream. Furthermore, with the responsibility of producers of AI as a component (within the meaning of product liability law), responsibility is extended even further upstream along the supply chain. This extension may even encompass providers of AI models, who supply foundational AI models or systems that are subsequently integrated into downstream products or services. Such an upstream allocation of liability reflects the principle that those who design and develop AI systems are often best positioned to identify and mitigate potential risks inherent in their products, and should therefore bear corresponding legal responsibility.

112 This upstream shift in responsibility has significant implications for the distribution of liability along the AI supply chain. While deployers remain subject to obligations under the AI Act and may incur liability for their own conduct in deploying AI systems, the product liability framework ensures that providers and component manufacturers cannot escape responsibility by contractually transferring risk downstream. This approach aligns with the broader policy objective of incentivising safety and quality assurance at the source of AI development, where technical expertise and the capacity to implement effective risk mitigation measures are typically concentrated.

113 In conclusion, the existing liability framework - consisting of the amended product liability and general national negligence liability - allows an implementation of our suggestions on tort liability as both - defectiveness of AI placed on the market by economic operators and negligence of AI providers and deployers - rely on breaches of duty. Our suggestions for certain monitoring obligations can be implemented as a part of these duties of care. However, this does not encompass our suggestions for strict liability in certain cases which is a legislative proposal.

II. Constant Monitoring with Human Oversight as a Part of Diligence in Liability Law (for Big AI)

114 We continue by examining how a system of constant monitoring with human oversight (as suggested in Section B.II.2) can be a part of due diligence under existing liability law and might be applied as part of due diligence duties within a framework of liability law. In this context, we analyse what legal consequences arise from not implementing the monitoring system at all versus implementing it and still causing damage.

1. First Case: No Implementation of the Monitoring System

115 In case AI developers do not implement the monitoring system described above at all, there is in our opinion a strong case that - unless alternative means to effectively address potential harm caused by AI are implemented at the output level - they act negligently and can, subject to other requirements for liability, be held liable. Scenarios of non-implementation (and adopting no alternative means to prevent AI harm) can therefore be considered a relatively clear and unproblematic case from a legal perspective. Assuming the injured party can show that the damage it suffered has been caused by the AI provided, under both fault-based and strict liability, the developer would be liable.

2. Second Case: Compliance with Implementation Obligation

116 While the first case is relatively easy to assess from a legal point of view, the second scenario poses a more complicated problem. Once deployers have implemented the constant monitoring system and a change in the misclassification rate, trend, or variance is detected, it is in principle still necessary to prove that the AI deployer has actually breached his duty of care. In fact, such a change may also have occurred through no fault of the latter. Hence, the mere fact that the system detected a change, does not necessarily constitute the deployer's breach of duty. However, in case the AI deployer fails to take any remedial action once a relevant change has been detected, such inaction despite knowledge of a negative change in performance, combined with

the possibility to act, would very likely have to be understood as a breach of the relevant duty of care - hence triggering the provider's liability. Given that the purpose of a constant monitoring system is primarily to detect change so that the AI operator can take action to prevent such change causing (further) damage, prompt action will be required and failure to do so will usually trigger liability.

117 While this does not resolve the issue of whether and when a change in AI performance as such constitutes a breach of a duty of care, liability triggered by inaction is likely to serve as a robust mechanism to minimise harm caused by AI. This approach is known from so-called "notice and takedown" obligations,[99] where operators can rely on safe harbours (i.e., no or limited liability) until they become aware of an infringement (e.g., illegal content) on their systems and are then liable from the moment they obtain knowledge if they do not act swiftly to remedy the infringement (e.g., remove content).

118 With regard to the outstanding issue of the detection of change as such being indicative of a breach of a duty of care, one needs to consider that random tests cannot normally prove a breach of duty per se. As shown before (Section C.I.1.a), a basic idea underpinning tort liability is that wrongful behaviour must be proven by the injured party. Showing a change in AI performance as such does not necessarily serve as evidence for breaching a duty of care: much depends on what caused that change, and whether the AI deployer could have prevented that change. Nevertheless, the inaction after a negative deviation from the required performance has been detected, can trigger liability.

119 The solution to address situations where a change in AI performance as such is the basis for a tort claim could consist of reversing the burden of proof. In that scenario, the AI deployer would have to prove that the change in performance does not reflect a breach of duty.[100] Such proof should normally be possible whenever the change is triggered by conditions outside the control of the AI provider - such as a shift in how for example users engage with an AI-based recommendation system. Even in cases where the AI deployer cannot exculpate itself (i.e., where the presumed breach of a duty of care stands), it would, however, still be up to the injured party to prove that the change occurred, as well as causality and damage. Alternatively, a strict liability regime could be implemented. In that case, the detected change in performance and the operator's inactivity would be sufficient as long as the injured party can prove the damage. The latter of these suggestions would, however, require a change in the existing (EU) law. However, under art. 10 sec. 4 of the new Product Liability Directive, the "defectiveness of the product or the causal link between its defectiveness and the damage, or both" can be presumed if "the claimant faces excessive difficulties, in particular due to technical or scientific complexity, in proving the defectiveness of the product or the causal link between its defectiveness and the damage, or both" (lit. a) and the claimant demonstrates that the causal link is "likely" (lit. b).

III. Implementation of a Cumulative Penalty System as Due Diligence

120 A cumulative penalty system as described above (see Section B.II.2.b) cannot be allocated to and executed within liability law. Setting up a penalty system contingent on legal assessments executed by third parties exceeds what is to be expected under conventional duties of care or due diligence. That aside, the penalties issued under the system as discussed above have punitive character, and do not correspond to the harm caused - whereas liability under tort law is governed by the general idea that those who cause harm have to restitute the situation that existed without the harmful event, or pay damages accordingly.

IV. Constant Monitoring as Part of an Early Warning Mechanism (for Big AI)

121 This Section considers existing approaches applied to online intermediaries, such as internet service providers (ISPs) and in particular (very) large platforms facilitating various forms of user interactions. As above, the analysis focuses on whether these approaches might provide for a suitable legal framework and/or useful implementation guidance for devising and operating a constant monitoring system with human oversight for Big AI, coupled with a cumulative penalty system, while adding some remarks on questions of liability and how they had been handled in relation to ISPs. Section C.IV.1 begins with a brief description of the conventional approaches adopted in the U.S. and the EU when online intermediaries emerged as key actors in the 1990s. It also introduces the more interventionist approach under the Digital Services Act (DSA) of the EU. Section C.IV.2 then looks at the specific legal tools available under the DSA, with a focus on measures applied to the operations of very large online platforms and search engines (VLOPSEs)

99 See e.g., for hosting providers, such as Internet Service Providers (ISPs), acc. to art. 16 DSA. A more detailed explanation of the concept will be given in Section C.IV.1.

100 Bundesgerichtshof [BGH] [Federal Court of Justice], Urteil vom 8. Dezember 1992 [judgement on Dec. 8, 1992], VI ZR 24/92 ("Mineralwasser I").

such as risk assessment, risk mitigation, independent audits and other oversight and enforcement mechanisms. Since the legal framework for online intermediaries has its foundations in exemptions from legal responsibility (in particular tort liability) of such intermediaries, but (under the DSA) has been extended to cover various means of direct regulation, it is presented here as a sui generis approach - somewhat in-between indirect regulation via liability law (see Sections C.I and C.II above), and direct regulation (see Section C.V below).

1. Introduction to Platform Regulation

122 When the Internet and other means of communication and data exchange through global digital networks gained commercial importance in the 1990s, the decentralised network structure demanded tools that would help users to navigate the increasingly complex network. ISPs and other service providers that mediated interaction among users emerged as essential facilities, serving a public function in terms of access to and sharing of information and other content, offering new dimensions for free speech and the ability to communicate across the globe. These providers in principle operate as private actors, not least because the rise of the (commercial) Internet coincided with globalising the 'Washington Consensus' of neo-liberal economic policies, including in the form of deregulation and privatisation of telecommunication services and related public utilities. As private actors, for a long time and up until recently, these intermediaries had not been subject to significant public law regulation - but of course remained potentially liable under private law vis-a-vis their own users and third parties. Due to their 'bottleneck' function as an essential facility for users' interactions online, ISPs ran the risk of being held liable for the conduct of their users - such as copyright infringements, defamation or hate speech. In order not to unduly impede ISPs (and in particular the development of network technologies used in their operations), the U.S. and then the EU adopted legal solutions for intermediary liability.[101] These solutions aimed to strike a balance between (1) utilising the ISPs' ability to prevent infringements and harm spreading through a network with (2) facilitating innovation in platform and other network technology.

123 In fact, convergent developments in this area can be observed across the globe. Notably, over the past decades a variety of different countries have adopted specialised sets of rules to govern the liability of intermediaries whose services are used by third parties to commit an infringement on the Internet.[102] These often take the form of so-called 'safe harbours', i.e., conditional liability exemptions that shield intermediaries from (some) legal responsibilities that they might otherwise incur, for example for actions that constitute participation in IP infringement or dissemination of hate speech. Notable examples of such rules include sec. 512 of the US's Digital Millennium Copyright Act (DMCA) and Arts. 13-14 of the EU's E-Commerce Directive.[103] While difficult to generalise, the most important common element of these safe harbours can be described by the 'notice and take-down' formula: With regard to content that is not their own (but of their users), ISPs have to act swiftly in removing or blocking access to illegal content - once they become aware of its illegal nature, usually once being notified. However, while these regimes display obvious similarities, significant differences also exist between them. So, for example, while s. 512 DMCA immunises intermediaries in the provision of four different kinds of services (mere conduit in s. 512(a), caching in s. 512(b), hosting in s. 512(c) and search services in s. 512(d)), the EU's E-Commerce Directive only protects intermediaries in the provision of the first three of these (mere conduit in art. 12, caching in art. 13 and hosting in art. 14). The DMCA is additionally limited to copyright infringement only, while the E-Commerce Directive applies horizontally to liability in a broad array of areas of civil and criminal law.

124 Differences also exist with regard to the reach of the safeguards provided. As opposed to their US equivalents, in the EU the safe harbours immunise intermediaries against liability only in the strict sense, i.e., for monetary compensation. All three of the EU safe harbours contain express permissions in their final paragraphs regarding the imposition of any kind of injunctive order on the providers of information society services by 'courts and administrative authorities' to 'terminate or prevent an infringement'.[104] Moreover, specifically in the area

101 See generally, Riordan, *The Liability of Online Intermediaries*, Oxford University Press (Sep. 7, 2016); Angelopoulos, *European Intermediary Liability in Copyright - A Tort-Based Analysis*, University of Amsterdam (Apr. 22, 2016).

102 See Seng, *Comparative Analysis of the National Approaches to the Liability of Internet Intermediaries* (2010), https://www.wipo.int/export/sites/www/copyright/en/docs/liability_of_internet_intermediaries.pdf (last accessed Feb. 28, 2024); and a more recent overview by Schmon, Pedersen, *Platform Liability Trends Around the World* (May 19, 2022), https://www.eff.org/deeplinks/2022/05/platform-liability-trends-around-globe-safe-harbors-increased-responsibility (last accessed Feb. 28, 2024).

103 Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), 2000 O.J. (L 178) 1, hereinafter: E-Commerce Directive.

104 Art. 12 sec. 3, 13 sec. 2, 14 sec. 3 and Recital 45

of copyright, EU law goes one step beyond this passive limitation: Recital 59 of the Infosoc Directive[105] observes that, in many cases, intermediaries are best placed to bring infringing activities occurring on their digital premises to an end. Art. 8 sec. 3 of the Infosoc Directive then explicitly instructs Member States to 'ensure that rightsholders are in a position to apply for an injunction against intermediaries whose services are used by a third party to infringe a copyright or related right'. Art. 11 of the 2004 Enforcement Directive[106] has subsequently expanded this obligation to cover all intellectual property rights.[107]

125 At the same time, there are limits on the injunctive orders as well as the scope of duties of care against intermediaries under EU law: a significant limitation on the permissible scope of both is imposed by art. 15 of the E-Commerce Directive, which prohibits the imposition of general obligations on service providers, in the provision of services protected by the three safe harbours, to monitor the information which they transmit or store or to actively seek facts or circumstances indicating illegal activity. The law of fundamental- and human rights also imposes limitations in this regard. The implications of this are considered in the next section when the more specific commitments now imposed under the DSA are reviewed for their suitability as a legal framework for an AI monitoring mechanism. As a successor to indirect platform regulation via the safe harbours in the E-Commerce Directive, the DSA takes a significantly more comprehensive and 'regulatory' approach.[108] In addition to provisions on ISP liability, the bulk of DSA rules concern 'Due Diligence Obligations for a Transparent and Safe Online Environment' (chapter III, arts. 11-48 of the DSA), including provisions on platform terms and conditions, transparency reporting, notice and action mechanisms, reasoned decisions, complaint handling and out-of-court settlements, trusted flaggers, recommender system transparency, protection of minors, as well as a range of commitments for specific types of online service providers.[109] While the next section can only address some of the elements of this approach when potentially relevant to AI monitoring, the DSA overall presents itself as the EU's complex regulatory response to increasing platform power, and a more diversified field of user - platform interactions.

126 In sum, ISPs or platform regulation emerged in the early days of the Internet primarily as a safe harbour mechanism whereby intermediaries can benefit from conditional exemptions from (civil) liability. Arguably, these exemptions created important legal foundations for platforms' business models, and to thrive economically. While there is no global set of harmonised rules (neither on indirect liability, nor on the safe harbours to avoid it), national approaches have - by and large and with some important differences - converged around a 'notice and take-down' approach. Such an approach eventually aims to balance the potential for the ISP to take effective action vis-a-vis its users when they commit allegedly illegal acts with the need not to impede the development of new technologies, also taking into account the (fundamental) rights of users, ISPs and affected third parties. We argue that, despite important differences in the factual settings at issue, a similar notion of balancing is at the heart of the AI monitoring system proposed here - but with a somewhat open question who is best placed to undertake this balancing, in particular in order to avoid potentially significant or even irreparable harm for more diffuse interests at a societal level. In addition, the increasing exposure of individuals (users and those directly or indirectly affected), groups, and societies to the often global reach of the most popular platforms corresponds to the exposures discussed in Section B.II to the reach of Big

E-Commerce Directive. Art. 18 E-Commerce Directive also requires that Member States 'ensure that court actions available under national law concerning information society services' activities allow for the rapid adoption of measures, including interim injunctions, designed to terminate any alleged infringement and to prevent any further impairment of the interests involved.' Cf. s. Digital Millennium Copyright Act (DMCA) § 512(j), 17 U.S.C.(1998) which does offer some limited possibilities of injunctive relief against intermediaries protected by one of the US safe harbours.

105 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, 2001 O.J. (L 167) 10, hereinafter Infosoc Directive.

106 Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights, 2004 O.J. (L 157) 45.

107 On the specific issue of injunctions against ISPs, see generally Husovec, *Injunctions against Intermediaries in the European Union: Accountable but Not Liable?*, Cambridge University Press (2017), https://doi.org/10.1017/9781108227421 (last accessed Feb. 28, 2024).

108 For a general discussion on the (European) 'new' regulatory approach to platforms, see Wagner et al., *Taming the giants: The DMA/DSA package*, 58 Common Market L.Rev. 987 (2021), https://doi.org/10.54648/cola2021065 (last accessed Feb. 28, 2024) and Iliopoulou-Penot, *Governing the 'Governors' of online speech: a review of the Digital Services Act's first steps in regulating social media platforms*, Yearbook of European law, 2025-10, https://doi.org/10.1093/yel/yeaf006 (last accessed Jan. 5, 2026).

109 On the DSA see generally, Husovec, *Principles of the Digital Services Act*, Oxford University Press (2024).

AI - so that arguments for regulatory intervention to prevent harm also align. With this in mind, the next section considers the regulatory tools of the DSA, especially those governing platforms with an equivalent harm-potential as Big AI.

2. Implementing a Constant Monitoring System with Human Oversight Based on the DSA

127 This section looks at the elements of the DSA's regulatory response to increasing platform power with a focus on those that may serve as a useful template for our proposed AI monitoring system with human oversight and a cumulative penalty mechanism, as described in detail in Section B.II.2.

128 To begin with, one might consider the foundational approach of 'safe harbour' liability exemptions for ISPs to be antithetical to the idea of devising a legal framework under which (Big) AI operators are responsible for preventing harm, or accountable for reparations should harm occur. And since the DSA retains the notion of safe harbours, including the 'no general obligation to monitor' principle,[110] its approach might appear ill-suited to guide the legal implementation of an AI monitoring system. However, the basic liability rules in Chapter II of the DSA concern responsibility not for own conduct, but that of *third* parties (*user* content and *acts of users)*; and they are driven by the broader context described in Section C.IV.1. While this can be distinguished from our concern over harm caused by AI, the due diligence obligations in Chapter III of the DSA appear as a much closer fit - in particular those related to 'very large online platforms' and 'very large search engines' (VLOPSEs) where the DSA takes a more regulatory, interventionist approach. Given that the deployment of Big AI might frequently take place via systems operated or supported by VLOPSEs and since the systemic risks identified with Big AI significantly overlap with those posed by VLOPSEs, the DSA regulatory approach to VLOPSEs can serve as a useful template.[111] This section hence focuses on these additional commitments for VLOPSEs in arts. 33 - 43 DSA, and the corresponding enforcement provisions in Chapter IV of the DSA.

a.) General Due Diligence Commitments for ISPs

129 Before considering in more detail the commitments for VLOPSEs pertaining to risk assessment and risk mitigation, it is worth highlighting some of the DSA provisions not specific to VLOPSEs, but with a potential guiding function for regulating harm caused by (Big) AI. To begin with, the DSA emphasises in various provisions the human rights implications of ISPs, online platforms and VLOPSEs in particular. Art. 1 DSA sets out its principal aim as, inter alia, creating a 'trusted online environment that facilitates innovation and in which fundamental rights enshrined in the Charter, including the principle of consumer protection, are effectively protected'. In relation to content moderation (including by means of AI) for example, art. 14.4 obliges all ISPs to 'act in a diligent, objective and proportionate manner' (....) 'with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter.' Under the risk assessment and mitigation obligations for VLOPSEs (discussed in more detail below), one of the four systemic risks which VLOPSEs have to assess and mitigate concerns 'any actual or foreseeable negative effects for the exercise of fundamental rights'.[112] The corresponding recital 81 DSA explains that '[s]uch risks may arise, for example, in relation to the design of the algorithmic systems used by the very large online platform or by the very large online search engine'. Similar commitments to take account of the fundamental rights of users can be found in arts. 36 and 38 DSA.[113]

110 See chapter II, art. 4-6, 8 DSA.

111 See Section B.II above on a discussion of the systemic risks posed by Big AI. The systemic risks posed by ISPs and in particular by VLOPSEs are discussed in recitals 1-3, 40, 47, 75-76, 79-90 DSA with recital 76 DSA setting out the basic idea that due to the potential for VLOPSEs to 'cause societal risks, different in scope and impact from those caused by smaller platforms', they 'should therefore bear the highest standard of due diligence obligations, proportionate to their societal impact'. For a similar approach, analysing DSA transparency obligations for 'high impact' AI, see Söderlund et al., *Regulating high-reach AI: On transparency directions in the Digital Services Act*, Internet Policy Review, 2024-01, Vol.13 (1), 1 - online at https://policyreview.info/articles/analysis/regulating-high-reach-ai (last accessed Jan. 2, 2026).

112 Art. 34.1 (b) DSA. The provision continues to list 'in particular the fundamental rights to human dignity enshrined in art. 1 of the Charter, to respect for private and family life enshrined in art. 7 of the Charter, to the protection of personal data enshrined in art. 8 of the Charter, to freedom of expression and information, including the freedom and pluralism of the media, enshrined in art. 11 of the Charter, to non-discrimination enshrined in art. 21 of the Charter, to respect for the rights of the child enshrined in art. 24 of the Charter and to a high-level of consumer protection enshrined in art. 38 of the Charter.' See also art. 35 sec. 1 DSA whereby the mitigating measures adopted to address the systemic risks need to specifically focus on the impact on fundamental rights.

113 See also the further elaborations on the diversity of

130 While the approach taken in the DSA involves broader questions about the extent to which ISPs as private actors should be human rights duty bearers which we do not discuss here,[114] the principal analogy drawn between ISPs, in particular VLOPSEs, and Big AI in terms of the respective potential to cause societal (if not global) harm appears prima facie defensible: Given (1) the direct link set out in recital 81 DSA (whereby one of the risks to the enjoyment of fundamental rights is posed by the design of AI systems employed by VLOPSEs); and (2) the similarity of circumstances between ISPs, in particular VLOPSEs, and Big AI in terms of individual and societal impact, it is worth considering whether the constant monitoring system for Big AI should equally include an assessment of the impact Big AI has on fundamental rights. This would primarily focus on the rights of those directly impacted by Big AI - the end users of AI systems, as well as those in relation to whom an AI system is applied. Ideally, the human-led monitoring process we propose would also include the impact on fundamental rights of those indirectly affected - that is, to the extent such an indirect impact is reasonably foreseeable, or to the extent information about such an indirect impact becomes available. However, as the monitoring system we advocate focuses on reviewing random samples of AI output/decisions (without those being tied to an individual complaint or dispute from which to gather further information about relevant human rights impact), the review will have to focus on human rights implications that are apparent from the AI output/decision (such as freedom of speech when legal content is filtered, or the rights of children if inappropriate content is not blocked). As discussed in Section B.II.4 above, the constant monitoring system has to focus on more 'tangible' legal interests, the impact on which can actually be assessed in full (i.e., without the need for further contextual information) by the CA on the basis of the samples provided.

131 Before moving on to the specific risk assessment and mitigation obligations for VLOPSEs, one should also mention that the DSA includes rules on transparency reporting on content moderation (including the extent to which automation has been relied on, and with what result).[115] Since the reports on automated content moderation have to include 'indicators of the accuracy and the possible rate of error of the automated means used', adhering to such reporting obligations implies a constant monitoring and review of AI-based content moderation tools and hence aligns with the basic idea of the constant monitoring system we propose. In fact, complying with the reporting commitments on AI-accuracy and error rates for content moderation might well necessitate a monitoring system equivalent to the one proposed in Section B.II.2. Another basic element of our proposal that finds support in the DSA is the concept of differentiating commitments depending on the size and type of the operator.[116] Furthermore, the idea of certifying dispute settlement providers for disputes between ISPs and users (with the possibility of revocation);[117] and the notion of 'trusted flaggers' (which serve to flag illegal content)[118] are elements of the DSA that can offer useful input for

contexts for and roles of fundamental rights especially in recitals 47, 51, 54, 63, 79, 81, 86, 109, and 153 DSA. On fundamental rights under the DSA, see Frosio et al., *Taking fundamental rights seriously in the Digital Services Act's platform liability regime. Eur Law J.* 2023; 29(1-2): 31-77. https://doi.org/10.1111/eulj.12475 (last accessed Jan. 5, 2026).

114 See on this United Nations, *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*, Final report to the Human Rights Council (A/HRC/17/31, Geneva, 2011). The UN Guiding Principles have been developed by J Ruggie, then Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.

115 See art. 15 sec. 1 (c) and in particular (e) DSA whereby reporting has to cover 'any use made of automated means for the purpose of content moderation, including a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used in fulfilling those purposes, and any safeguards applied', as well as art. 24 DSA for online platforms (concerning further reporting commitment on internal complaints, disputes, and abuse-countermeasures), and art. 42 DSA for VLOPSEs. See generally Enarsson, *Transparent, Rapid and Contextualised? Comparing Content Moderation Requirements in Emerging EU Regulation*, European journal of law and technology, 2025-05, Vol. 16 (2), 1 & Söderlund et al., as note 118.

116 See arts. 4-6, 19, 20-28, 29, 33 DSA (including an exemption under art. 19 from many duties for micro and small enterprises with less than 50 employees and no more than EUR 10 million turnover, as well as detailed rules in art. 33 on how to determine when an online platform is 'large enough' to constitute a 'very large online platform', triggering the corresponding significant additional commitments).

117 See art. 21 DSA setting out the principal standards against which certifications are to be issued, the requirements and procedures for dispute settlement providers to adhere to when offering their services, and related reporting obligations. In case it is not feasible for the monitoring of AI under our proposal to be not conducted by state authorities, similar certification principles for private providers to conduct the monitoring may be considered.

118 In particular, the selection criteria for trusted flaggers (such as expertise, independence, diligence and objectivity), as well as the reporting standards of trusted flaggers could serve as a basis for the experts relied on in our system of human monitoring for AI, in particular if such monitoring had to be conducted by those other than state authorities; see arts. 16 sec. 1, 22 sec. 2 and 3 DSA.

the implementation of the proposed AI monitoring system.

b.) Risk Assessment and Mitigation

132 The most relevant aspects of the due diligence commitments under Chapter III of the DSA, however, relate to risk assessment and mitigation by VLOPSEs. Recitals 75, 76 and 79 DSA explain the motivation for imposing these commitments: Due to the 'reach' of VLOPSEs[119] they 'may cause societal risks, different in scope and impact from those caused by smaller platforms'.[120] Given that VLOPSEs as private actors design and operate their services in order to maximise profit and/or optimise their business models, the societal concerns caused by their reach are externalised. Hence, 'effective regulation and enforcement is necessary in order to effectively identify and mitigate the risks and the societal and economic harm that may arise'.[121] VLOPSEs 'should therefore bear the highest standard of due diligence obligations, proportionate to their societal impact'.[122] The DSA assumes that once a platform exceeds a given number of active users (currently set at 45 million per month)[123] the systemic risks posed 'may have a disproportionate impact in the Union'. For these reasons, the risk assessment and mitigation commitments are considered 'necessary' in order to be able to address public policy concerns resulting from the potential impact of VLOPSEs, there being - according to the EU - 'no alternative and less restrictive measures that would effectively achieve the same result.'[124] Based on our definition of Big AI, the analogous problems of (1) externalising public policy concerns, as well as (2) the far-reaching and systemic impact on societies,[125] offer a strong argument for adopting at least an equivalent approach to Big AI operators. As shown below, the specific DSA commitments for VLOPSEs align significantly with our suggestions for a constant monitoring mechanism for Big AI.

133 According to art. 34 sec. 1 DSA, VLOPSEs are under an obligation to 'diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services'. These risk assessments have to be carried out at least once every year, and 'in any event prior to deploying functionalities that are likely to have a critical impact on the risks identified pursuant to this Article'.[126] The VLOPSEs risk assessment must be 'specific to their services and proportionate to the systemic risks, taking into consideration their severity and probability'.[127] Art. 34 sec. 1 DSA then defines four particular categories of systemic risks VLOPSEs have to assess:

> (a) 'the dissemination of illegal content through their services';
>
> (b) 'any actual or foreseeable negative effects for the exercise of fundamental rights';
>
> (c) 'any actual or foreseeable negative effects on civic discourse and electoral processes, and public security'; and
>
> (d) 'any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being'.[128]

134 When assessing these risk categories, art. 34 sec. 2 DSA requires VLOPSEs to consider the

119 Recital 75 DSA describes such 'reach' in terms of facilitating public debate, economic transactions and the dissemination of information, opinions and ideas and in influencing how users obtain and communicate information online; and for search engines in terms of locating and making information retrievable online.

120 See recital 76 DSA. Recital 79 DSA describes these risks as the potential for VLOPSEs to be used in a way that 'strongly influences safety online, the shaping of public opinion and discourse, as well as online trade'. On VLOPSEs generally, see Moravcová, *Impact of the DS Regulation on Very Large Online Platforms*, Central European journal of comparative law (Online), 2023-12, Vol.4 (2), 163.

121 See recital 79 DSA.

122 See *id.* recital 76 DSA.

123 Art. 33 sec. 1 DSA - a figure that, according to recital 76 DSA, corresponds to 10% of the current population in the EU. Art. 33 DSA further sets out details on how to adjust this number, depending on the EU population, and how to determine whether a platform or search engine falls under this category. For a list of those actors that currently fall under this category, see https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses.

124 Recital 76 DSA.

125 See Section B.II above.

126 Art. 34 sec. 1 DSA. In contrast to this, the constant monitoring system proposed here would involve much more frequent, although less in-depth scrutiny of AI outputs (hence capturing primarily direct observable harm very swiftly - whereas full fledged audits under the DSA are better suited for diffuse harms).

127 Recital 79 DSA further explains that in determining the significance of potential negative effects and impacts, providers 'could assess whether the potential negative impact can affect a large number of persons, its potential irreversibility, or how difficult it is to remedy and restore the situation prevailing prior to the potential impact.'

128 See art. 34 sec. 1 (a)-(d) DSA. On point (b), see also the discussion above (incl. notes 118-120) concerning the link between human rights, VLOPSEs, Big AI impact and risk assessment.

impact of factors such as (a) 'the design of their recommender systems and any other relevant algorithmic system'; (b) 'their content moderation systems'; (c) 'the applicable terms and conditions and their enforcement'; (d) 'systems for selecting and presenting advertisements'; and (e) 'data related practices of the provider'. Assessment of the systemic risk categories (a)-(d) under art. 34 sec. 1 DSA also has to consider the potential for 'intentional manipulation of [the VLOPSEs'] service, including by inauthentic use or automated exploitation of the service', as well as the 'amplification and potentially rapid and wide dissemination of illegal content'.[129]

135 The analogy to the potential risks posed by Big AI is apparent: First, the four categories of systemic risks identified in art. 34 sec. 1 DSA equally apply to the deployment of Big AI - not least since scale, speed and severity of any potential harm caused by VLOPSEs in these categories will usually result from VLOPSEs' own reliance on AI and other means of automation when providing their services. In fact, four of the five factors in art. 34 sec. 2 (a) - (e) DSA relate to the use of AI and other automatic means in supplying VLOPSEs services, again illustrating the impact Big AI can have on individuals, groups and societies. Finally, the potential for abuse as well as the amplification of impact referred to in art. 34 sec. 2 DSA is particularly acute when AI is deployed: recital 84 DSA explicitly refers to VLOPSEs duty to consider how 'algorithmic amplification of information contributes to the systemic risks'. Given this overlap, the risk assessment obligations under art. 34 DSA can serve as a useful template for a legal framework applicable to risk assessments related to the deployment of Big AI. In fact, given that many situations where VLOPSEs have to assess (and then mitigate) risk involved in the deployment of AI, our model for a constant monitoring system in turn can, subject to some important modifications,[130] serve as a template for implementing the commitments under art. 34 DSA. At the same time, the nature of the constant output monitoring mechanism we propose does not lend itself to a thorough analysis of latent and potentially remote risks to less concretely formulated legal interests, as reflected in some human rights and in broader societal concerns for democratic, representative and non-discriminatory governance structures.

136 The principal distinction between the risk assessment and mitigation framework of the DSA and our proposal for constant monitoring of Big AI then primarily lies in the greater autonomy provided to the private operator under the DSA: our model suggests constant monitoring and oversight *by an expert third party* (our 'check authority', CA) to prevent harm from materialising - rather than leaving risk handling primarily in the hands of the private actor. Our monitoring proposal further involves more frequent, constant control that is meant to detect harm quickly and early - rather than the DSA risk assessment, in principle to be conducted annually. At the same time, our proposal is narrower in terms of *limiting the risk monitored* to legal interests whose harm can be observed from the samples available to the CA. Despite these distinctions, the DSA also foresees further oversight and control. Under art. 35 DSA, VLOPSEs have to 'put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to art. 34 DSA, with particular consideration to the impacts of such measures on fundamental rights'.[131] Art. 35 DSA leaves the specific mitigatory response to the platform - but suggests this may consist, inter alia, in adapting VLOPSEs' (1) online interfaces, design, features of functioning; (2) their content moderation systems; or in (3) 'testing and adapting their algorithmic systems, including their recommender systems'. To the extent these measures would involve automated controls or tweaks in VLOPSE's AI systems, they would be subject to the same criticism we discussed regarding relying on an 'AI checking an AI'. Checking compliance with these due diligence mitigation obligations for VLOPSEs then is implemented through reporting duties, independent audits, data access and scrutiny, as well as the enforcement powers for national authorities (acting as Digital Service Coordinators, DSCs), the European Board for Digital Services (EBDS) and the European Commission (EC), including the

129 See art. 34 sec. 2 DSA. On addressing systemic risks under the DSA see also Kaesling, Wolf, *Sustainability and Risk Management under the Digital Services Act: A Touchstone for the Interpretation of 'Systemic Risks'*, GRUR International, 2025-02, Vol.74 (2), 119; Liesenfeld, *The Legal Significance of Independent Research based on Article 40 DSA for the Management of Systemic Risks in the Digital Services Act*, European journal of risk regulation, 2025-03, Vol.16 (1), 184.

130 Since art. 34 DSA primarily foresees the risk assessment to be conducted by VLOPSEs (subject to review in accordance with arts. 37, 40, 49-50 DSA - discussed below), our proposal for a constant monitoring system by an external entity would require some important adaptation. Despite these principal differences VLOPSEs could 'test' their systems using a constant monitoring model as described in Section B.II.2.a, in particular as a precautionary pre-deployment measure to limit directly observable errors leading to potential harm.

131 Recital 86 DSA adds that these mitigation measures 'should be proportionate in light of the economic capacity of the provider of the very large online platform or of the very large online search engine and the need to avoid unnecessary restrictions on the use of their service, taking due account of potential negative effects on those fundamental rights. Those providers should give particular consideration to the impact on freedom of expression.'

possibility for significant and periodic penalties.[132] Since there is no room here to fully review these oversight and enforcement tools, the following summary merely highlights the main overlaps with our proposed constant monitoring system.

c.) Oversight and Enforcement Tools

137 Under art. 37 DSA, VLOPSEs 'shall be subject, at their own expense and at least once a year, to independent audits to assess compliance'. To enable such audits in an 'effective, efficient and timely manner' VLOPSEs have to provide independent auditors 'access to all relevant data and premises and by answering oral or written questions' - who then have to produce substantiated written audit reports.[133] In case the audit does not yield positive results, VLOPSEs 'shall take due account of the operational recommendations addressed to them with a view to take the necessary measures to implement them'.[134] Next to being subject to annual independent audits, VLOPSEs have to adopt response mechanisms in case of crisis (such as a serious threat to public security or public health in the EU)[135] and generally have to offer access to data (including data that 'explain[s] the design, the logic, the functioning and the testing of their algorithmic systems')[136] to the extent 'necessary to monitor and assess compliance with this Regulation'.[137] Data access has to be provided to the national DSC, or the EU Commission. The former function as the default oversight authority and might serve as a template for setting up a 'check authority' (CA) under our proposed monitoring system for Big AI.[138] The investigatory and enforcement powers of the DSC include powers to set fines and periodic penalty payments, escalating up to temporary access restrictions to platform services, and to adopt interim measures.[139] However, any of the enforcement measures taken must be 'effective, dissuasive and proportionate, having regard, in particular, to the nature, gravity, recurrence and duration of the infringement or suspected infringement to which those measures relate, as well as the economic, technical and operational capacity of the provider of the intermediary services concerned where relevant'.[140] In addition, these oversight and enforcement measures are subject to general safeguards under EU law.[141]

138 Subject to these safeguards, DSCs can issue penalties and fines against ISPs for breaching their due diligence obligations under Chapter III of the DSA. Under art. 52 DSA (and with some further qualifications for VLOPSEs under arts. 74-76 DSA), the maximum amount of fines for non-compliance 'shall be 6 % of the annual worldwide turnover of the provider of intermediary services concerned in the preceding financial year' and the maximum amount of a periodic penalty payment 'shall be 5 % of the average daily worldwide turnover or income of the provider of intermediary services concerned in the preceding financial year per day'.[142] The oversight

132 See generally Chapter IV, arts. 64-83 of the DSA, detailing the specific enforcement regime in relation to VLOPSEs. The EU has not shied away from relying on its enforcement powers under the DSA - for example in relation to TikTok and Meta (see: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503) as well as X (formerly Twitter - see: https://digital-strategy.ec.europa.eu/en/news/commission-fines-x-eu120-million-under-digital-services-act).

133 See art. 37 sec. 2 and 4 DSA. In addition, under art. 34 sec. 4 DSA VLOPSEs have to retain the supporting documents of their risk assessments for at least three years and, upon request, have to communicate these to the relevant enforcement agencies. Art. 37 sec.3 DSA further sets out requirements which aim to ensure the independence and impartiality of the auditors.

134 See art. 37 sec. 6 DSA.

135 See art. 36 and recital 91 DSA. As part of a crisis response mechanism, VLOPSEs have to 'apply specific, effective and proportionate measures to prevent, eliminate or limit any such contribution to the serious threat' (art. 36 sec.1 DSA). This is subject to oversight by enforcement authorities to ensure that the response measures are 'strictly necessary, justified and proportionate' (art. 36 sec. 3 DSA). In addition, there is a recommendation to draft crisis protocols under art. 48 DSA.

136 See art. 40 sec. 3 DSA. See also Liesenfeld, as note 136.

137 See art. 40 sec. 1 DSA.

138 See arts. 49, 50 DSA - whereby DSC have to 'perform their tasks (...) in an impartial, transparent and timely manner' and 'shall act with complete independence (...) free from any external influence, whether direct or indirect, and shall neither seek nor take instructions from any other public authority or any private party' (art. 50 sec. 2 DSA).

139 See art. 51 DSA.

140 See art. 51 sec. 5 DSA. In particular with regard to access restrictions (which might be an important sanctioning tool for Big AI as well) '[a]ny measure ordered shall be proportionate to the nature, gravity, recurrence and duration of the infringement, without unduly restricting access to lawful information by recipients of the service concerned' (art. 51 sec. 3 DSA).

141 See art. 51 sec. 6 DSA whereby 'any exercise of those powers is subject to adequate safeguards laid down in the applicable national law in compliance with the Charter and with the general principles of Union law. In particular, those measures shall only be taken in accordance with the right to respect for private life and the rights of defence, including the rights to be heard and of access to the file, and subject to the right to an effective judicial remedy of all affected parties'.

142 See also the specific powers to issue fines and periodic penalty payments to VLOPSEs under arts. 74, 76 DSA, discussed below. On 5 December 2025, the US-based platform X (formerly Twitter) had been fined a sum of 120

is further strengthened by an individual's right to lodge a complaint against an ISP (including of course VLOPSEs) with the DSC.[143] Somewhat related to such individual complaints, art. 54 DSA refers to national negligence liability rules whereby '[r]ecipients of the service shall have the right to seek, in accordance with Union and national law, compensation from providers of intermediary services, in respect of any damage or loss suffered due to an infringement by those providers of their obligations under this Regulation.' This suggests that a breach of the user-related obligations under the DSA also serves as basis for a tort claim under national liability law,[144] hence creating a similar parallelity of regulatory and liability frameworks as we have discussed above.

139 Finally, it is worth noting that under arts. 65-83 DSA, a detailed supervisory framework specific to VLOPSEs, operated by the EU Commission, exists. This framework aligns with the general implementation monitoring and enforcement powers discussed above, and - in addition - gives the EU Commission the right to:

- Initiate an investigation - *ex officio* or based on a request from a DSC - with a view to determining a VLOPSE's compliance with the DSA (arts. 65-66 DSA);
- Request information from the VLOPSE concerned, and/or any third parties who might be in the possession of relevant information (art. 67 DSA);
- Conduct investigations, including by means of access to and inspection of premises, books, data and other material of providers - if need be under reliance on coercive measures (arts. 68-69 DSA);
- Adopt any necessary interim measures in case of urgency due to the risk posed by the VLOPSE under investigation (art. 70 DSA);
- Make binding any commitments offered by the VLOPSE to ensure compliance with the DSA (art. 71 DSA);
- Take all necessary actions to monitor VLOPSE's compliance with the DSA, including requiring the retention of data or documents by the VLOPSE, and/or appointing independent auditors (art. 72 DSA);
- As part of a decision on non-compliance (art. 73 DSA), issue significant fines of up to 6% of the annual global turnover of the VLOPSE - but only in case of intentional or negligent breaches of the DSA (art. 74 DSA);[145]
- Engage in enhanced compliance supervision, such as demanding the VLOPSE to 'draw up an action plan setting out the necessary measures which are sufficient to terminate or remedy the infringement' (art. 75 DSA); and
- Issue 'periodic penalty payments not exceeding 5% of the average daily income or worldwide annual turnover in the preceding financial year per day', including to compel compliance with information requests, interim measures, and final decisions (art. 76 DSA).

140 All these supervisory and enforcement measures

million Euros for breaches of its transparency obligations under the DSA - see: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934.

143 See art. 53 DSA.

144 The further requirements of liability arguably derive from domestic law - see the discussion by Benjamin Raue in Hofmann/Raue (eds.), Digital Services Act: DSA, Article 54, para. 6-10.

145 In addition to fines as part of a decision on non-compliance, fines of up to 1% of the global annual turnover can be issued in cases of negligent or intentional non-compliance with the Commission's investigation (such as providing late, incomplete or misleading information); see *id.* art. 74 sec. 2 DSA.

at the hand of the EU Commission are, of course, subject to relevant limitation periods (arts. 77 and 78 DSA) as well as general due process guarantees (art. 79 DSA). The decisions of the EU Commission on measures adopted vis-a-vis VLOPSEs are to be made public (art. 80 DSA) and subject to review by the Court of Justice of the European Union (CJEU) in accordance with art. 81 DSA.

d.) Commonalities and Differences

141 In sum then, the DSA regulatory framework with significantly enhanced due diligence commitments, in particular for VLOPSEs, offers an interesting template for many aspects of our proposed monitoring system with human oversight and cumulative penalty system. While the DSA's approach to VLOPSEs does not exactly match our proposal, it reinforces the basic idea of preventing harm by a tailored mechanism of risk assessment and mitigation, further backed by monitoring, enforcement and, if necessary, sanctions. At the same time, the essential difference of internal (DSA) vs external (our proposal) monitoring makes sense in relation to the type of risks covered: we argue that external oversight based on samples of AI output is a safer, and hence preferable, way of addressing the risks which can actually be checked by a constant monitoring system - namely those where the harm materialises in relation to clearly defined and well understood legal interests. For the more remote or indirect risks to more diffuse, less 'tangible' legal interests (for example in relation to societal goals and governing principles), our constant monitoring based on samples will often not work well in detecting relevant harm merely on the basis of the facts observable from the samples. For these cases, the due diligence risk assessment and mitigation tools of the DSA might well serve as the, for now, best available option. A lot here will depend on factors such as: (1) how serious the VLOPSE takes its commitments to assess and mitigate systemic risks posed to human rights and other fundamental societal interests; (2) how rigorous external control by the DSC and the EU Commission is exercised; and - perhaps most importantly - (3) how much of the weighing and balancing inherent in making decisions about addressing systemic risks to less tangible legal and societal interest is effectively left to the private actor. One hopes that enhanced regulation *of* platforms via the DSA does not lead to ever more aspects of our (digital) lives being effectively regulated *by* platforms.

142 Returning to the comparison between the DSA model and our proposal, more fundamentally still, the underlying motivation for regulatory intervention is essentially identical: the systemic risk posed to individuals (both users and others directly or indirectly affected), groups and societies through private operators of systems that have unprecedented reach and their potential for all kinds of harm. In fact, many facets of this risk result from the common denominator of VLOPSEs and Big AI operators: the reliance on automated output that is scalable, fast and usually global in reach. VLOPSEs will frequently rely on Big AI systems in offering their services. For this reason, the regulation of VLOPSEs effectively often includes regulating Big AI (in particular to the extent VLOPSES deploy AI tools in supplying their services in the EU), along the lines of the DSA's internalised risk assessment and mitigation model with some degree of external oversight. At the same time, the DSA's approach still retains aspects of liability law (primarily in the form of the safe harbours discussed in Section C.IV.1) and hence needs to be distinguished from the direct regulatory approach to AI discussed in the next section.

V. Constant Monitoring Obligations as a Part of Direct Regulation (for Big AI)

143 We continue by examining how constant monitoring with human oversight can be a part of existing and upcoming regulatory law. As before, we start by giving a short overview of the basic concept of regulatory law. Thereafter, we will analyse if and where both random testing and the penalty system can be allocated in regulatory law.

1. Introduction to Regulatory Law

144 In contrast to liability law, regulatory law is a form of direct up front (ex ante) regulation which aims at preventing damage before it is caused. With regard to products, the main idea consists of the prevention of damage by setting rules for certification and accreditation. Regulatory law establishes obligations to be obeyed before placing an AI system on the market. In addition to that, it sets out continuous obligations which must be followed during the whole lifecycle of a product after it has been placed on the market.

145 Regulatory law is considered a direct form of regulation because its rules address the regulatee directly in a way that obligates the regulatee to act in a particular way, and/or not to commit certain acts; whereas liability rules function more indirectly by incentivising a certain behaviour in order to avoid liability.

2. Constant Monitoring with Human Oversight Under Existing Regulatory Law

146 Using EU law as principal legal reference point (not least because of the comparably more rigid approach to regulatory intervention in the EU, in contrast to an often more *laissez-faire* approach in the US), direct regulation is applied in many areas. Product safety, for example, is regulated by the "New Legislative Framework" (NLF) which sets out harmonised rules providers must comply with in order to be allowed to participate in the European internal market. In particular, the NLF contains a bouquet of Regulations and Directives laying down general rules for conformity assessment, CE marking, accreditation of conformity assessment bodies, and market surveillance.[146] Besides that, all sector-specific Regulations and Directives (i.e., Medical Devices Regulation,[147] Toy Safety Directive,[148] Construction Products Regulation,[149] etc.) are aligned with these general provisions. Hence, for each specific type of product, the same general rules concerning the conformity assessment, CE marking, and market surveillance apply. If AI systems are thus built into products whose categories are already covered by the NLF regulations, operators must comply with the NLF requirements under existing law if they wish to place their AI-driven products on the EU market.

a.) Constant Monitoring with Human Oversight as Part of Market Surveillance

147 Constant monitoring with human oversight could be implemented as part of the market surveillance laid down in the Market Surveillance Regulation (MSR).[150] As part of the NLF, the MSR aims at guaranteeing safe products by obliging operators to observe their products continuously after the latter have been placed on the market.

148 Pursuant to art. 11 sec. 3 MSR, market surveillance authorities shall perform appropriate checks on the characteristics of products on an adequate scale, by means of documentary checks and, where appropriate, physical and laboratory checks based on adequate samples. In addition to that, in deciding on which checks to perform, on which types of products and on what scale, market surveillance authorities shall follow a risk-based approach, taking into account possible hazards, the economic operator's record of non-compliance (as suggested above in Section B.II.2.b), and further factors. Hence, constant monitoring with human oversight could be implemented on the basis of the MSR.

b.) Constant Monitoring with Human Oversight as Part of Technical Norms and Standards

149 Moreover, the NLF is characterised by its specific regulatory structure: While both general and sector-specific Regulations and Directives such as the MSR and the Medical Devices Regulation set out general requirements for the respective products, concrete instructions for action are provided by technical norms and standards (provided by standardisation organisations, such as the ISO, CEN, CENELEC, and ETSI).

150 While the provisions of Regulations and Directives generally are binding (in that they set out mandatory legal commitments to be complied with), compliance with technical norms and standards is voluntary. However, compliance with such norms and standards triggers a rebuttable presumption that manufacturers comply with the mandatory legal requirements (presumption of conformity). For this reason, and because the technical norms and standards are largely developed on an industry-driven basis, they are highly relevant in practice and are implemented by most companies.

151 Monitoring obligations are already set out in existing technical standards, such as ISO 9001,[151] which lays down general rules and objectives for quality management, particularly for measuring and analysing customer satisfaction and product performance. In addition to that, sector-specific

146 European Commission, *New legislative framework*, European Union, https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en (last accessed Feb. 29, 2024).

147 Regulation of the European Parliament and of the Council 2017/745 of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, 2017 O.J. (L 117) 1, hereinafter Medical Devices Regulation.

148 Directive of the European Parliament and of the Council 2009/48/EC of 18 June 2009 on the safety of toys, 2009 O.J. (L 170) 1.

149 Regulation of the European Parliament and of the Council 305/2011 of 9 March 2011 laying down harmonised conditions for the marketing of construction products and repealing Council Directive 89/106/EEC, 2011 O.J. (L 88) 5.

150 Regulation of the European Parliament and of the Council 2019/1020 of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, 2019 O.J. (L 169) 1, hereinafter MSR.

151 International Organization for Standardization, *ISO 9001 Quality management systems – Requirements* (Sep. 2015), https://www.iso.org/standard/62085.html (last accessed Feb. 29, 2024).

standards (e.g., ISO 13485:2016[152] for medical devices) stipulate particular instructions for quality management (including monitoring obligations) which apply to products that belong to the respective sector. Hence, monitoring obligations are already part of technical norms and standards and compliance with these obligations triggers the rebuttable presumption that a product is compliant with EU law. As a result, the proposed constant monitoring system with human oversight could be implemented as part of (general or specific) technical standards for quality management as long as the AI system is embedded into a product which falls under the NLF scheme. A penalty system, however, cannot be implemented into a standard, since technical standards solely aim at providing concrete instruction for action. Thus our proposal for penalty systems (outlined above in Section B.II.2.b) goes beyond what is possible with standards.

152 At the time of writing, there are no specific standards for the quality management of AI systems. However, work is currently being done on such a standard[153] by the technical committee ISO/IEC JTC 1/SC 42. Thus, a constant monitoring system with human oversight which applies specifically to AI, could soon be implemented into a technical standard. The work of the responsible technical committee remains to be seen. The system we outlined above (see Section B.II.2) could serve as a basis for the committee's work on monitoring standards as requested by the EU COM in their standardisation request of May 2023 which has been accepted by the European standardisation organisation CEN and CENELEC.[154]

3. Constant Monitoring with Human Oversight Under Recently Enacted Direct Regulation

153 At the time of writing, Europe's first binding legal framework for AI has recently been finalised. After the Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (AI Act)[155] had been issued by the European Commission (EU COM) in April 2021, both the Council of the European Union (Council) and the European Parliament (EP) published their compromise amendments in November 2022[156] and June 2023,[157] all three legislative bodies entered trilogue negotiations in 2023. In December 2023, a political agreement was reached and a (provisional) consolidated version of the final text was issued in January 2024. Once the text was officially translated and both Council and European Parliament voted in favour of the new regulation, the regulation entered into force on 2 August 2024.[158] In general, according to art. 113, its rules will apply from 2 August 2026.

154 One central characteristic of the AI Act consists of its risk-based approach. This approach allows it to classify and regulate AI systems according to the risk they entail and the application field they are used in. The risk-based approach means that the only AI systems subject to strict requirements are those that are categorised as actually posing a high risk. Systems that pose a low or minimal risk, on the other hand, are subject to few or no obligations. A high risk is assumed where AI systems are used in areas already subject to third-party conformity assessments and certification procedures under

152 International Organization for Standardization, *ISO 13485 Quality management systems – Requirements for regulatory purposes* (Mar. 2016), https://www.iso.org/standard/59752.html (last accessed Feb. 29, 2024), which also contains monitoring obligations.

153 International Organization for Standardization, *ISO/IEC FDIS 42001 Information technology – Artificial intelligence Management System* (Dec. 2023), https://www.iso.org/standard/81230.html (last accessed Feb. 29, 2024).

154 European Commission, Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence (C(2023) 3215 final), https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en (last accessed Jan. 8, 2026).

155 European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM(2021) 206 final (Apr. 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (last accessed Feb. 29, 2024).

156 Council of the European Union, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules in artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts- General approach,* 14954/22 (Nov. 25, 2022), https://artificialintelligenceact.eu/wp-content/uploads/2022/12/AIA-%E2%80%93-CZ-%E2%80%93-General-Approach-25-Nov-22.pdf (last accessed Feb. 29, 2024).

157 European Parliament, *Amendments adopted by the European Parliament on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain legislative acts*, P_TA(2023)0236 (Jun. 14, 2023) https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf (last accessed Feb. 29, 2024).

158 Regulation of the European Parliament and of the Council 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

sector-specific regulation (e.g., medical devices) or where they are used in particularly sensitive areas (e.g., critical infrastructure) and pose significant risk.

155 Hence, if an AI system is implemented into a medical device which falls under the Medical Devices regulation and must undergo a third party conformity assessment (e.g., systems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act).an AI system which is implemented into MRI scannerssystems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act). using computer vision to detect anomalies) it is automatically considered high-risk under the AI Act (art. 6 sec. 1 and annex I sec. A no 11 of the AI Act). Regarding high-risk AI systems which are used in sensitive areas (systems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act).e.g., AI systems used to assess the creditworthinesssystems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act). of natural persons or AI systems used in the job application process), the operator, after testing the system, may argue that the system does not entail a high risk, as there is no significant risk in the specific use case scenario - for example, in the case of an AI system used to support human decision-making. And those AI systems which are deemed to entail a high risk according to the AI Act must comply with a number of requirements - among other things with the establishment of a quality management system (see arts. 8-21 of the AI Act).

156 A differentiation between Big AI and Small AI elaborated in Section A.V is not taken into consideration within the AI Act's risk classification. Hence, Small Team - Big AI, for example, would not fall under the high-risk group by default. The matrix illustrated in Section A.V could however be added to the annexes of the AI Act. So far, the annexes serve as substantiation of the legal text. Annexes I and III, for example, concretise which harmonised regulations and which areas of application are relevant for the risk classification. Small AI and Big AI could be added to annex III which sets out the areas or providers that pose a high risk. This way, the proposed system (matrix + monitoring obligation) could be adopted within the AI Act if Annex III was to be modified accordingly. The EU COM is entitled to amend annex III of the AI Act if necessary.

157 As mentioned above, for providers of high-risk AI systems the obligations pursuant to arts. 9 - 15 AI Act apply. Among these obligations are monitoring and testing of high-risk AI systems as part of the risk management pursuant to art. 9 sec. 2, 5, 6, 7 of the AI Act. In addition to that, high-risk AI systems must be tested in order to fulfil the data quality requirements set out in art. 10 AI Act. Continuous monitoring could also be implemented as part of the obligatory record-keeping (art. 12 sec. 2c AI Act) since monitoring goes hand in hand with keeping logs. Furthermore, continuous monitoring could be implemented as one, if not the most effective way to guarantee human oversight, which is required under art. 14 sec. 3 and 4. And finally, a constant monitoring system could be part of the required post market monitoring pursuant to art. 72 AI Act.

158 For deployers of high-risk AI systems, monitoring the system's operation will be obligatory pursuant to art. 26 sec. 5 AI Act. Hence, constant monitoring with human oversight would equally serve to implement obligations for AI deployers. For providers and third parties involved in the conformity assessment of high-risk AI systems, art. 43 AI Act as well as the annexes VI and VII of the AI Act or sector-specific regulation could be the linking point for the monitoring system proposed in this paper. Monitoring and testing could hence be part of the conformity assessment. Moreover, the cooperation between providers and notified authorities (equivalent to the Check Authority or "CA" discussed in Section B.II.2.a) is covered by art. 43 sec. 3 AI Act.

159 The specific process or monitoring loop would not be described in the AI Act itself but rather laid down by technical regulation (norms and standards, see above). The central provision establishing the rebuttable presumption of conformity in case harmonised standards are complied with, is art. 40 AI Act.

160 To summarise, it can be concluded that the AI Act provides various points of reference for the monitoring system developed in this paper. It can be subsumed under the obligations for high-risk AI systems and could therefore be the basis for technical norms and standards which concretise these obligations.

4. Implementing a Penalty System Under Direct Regulatory Law

161 Finally, the question arises whether in regulatory law, as opposed to liability law (see above), the penalty system described in Section B.II.2.b could be established. Based on general rule of law considerations, the fact that such a penalty system involves decisions imposing disadvantageous effects (fines and other penalties) on the regulatees then

constitutes an interference with the fundamental rights of the regulatees which requires, among other things, a legal basis.

162 One could consider introducing the penalty system as part of the market surveillance procedure pursuant to art. 72 ff. of the AI Act. Since market surveillance falls within the member states' competency, market surveillance authorities at national or federal level would be responsible. According to art. 74 sec. 1 of the AI Act, the general rules regarding market surveillance laid down in the Market Surveillance Regulation (MSR)[159] apply to the national competent authorities for AI market surveillance, too.

163 The MSR grants market surveillance authorities the power to restrict the rights of economic operators at various points. In particular, MSR provides for the following measures:

- Art. 14 stipulates the powers of market surveillance authorities, allowing them, inter alia, to impose penalties on economic operators (see art. 14 sec. 4 (i) MSR).
- Pursuant to art. 41 sec. 1 MSR, the Member States shall lay down the rules on penalties in case market surveillance obligations have been breached. These obligations can either result from the MSR or from sector-specific legislation listed in annex II. The Member States have the responsibility to take all measures necessary to ensure that those penalties are implemented in accordance with national law.
- According to art. 11 sec. 1 (b) MSR, market surveillance authorities are allowed to take appropriate and proportionate measures where the economic operator fails to comply with Union legislation and to take corrective action afterwards.

164 Hence, the penalty system introduced above could be implemented as part of the market surveillance according to the AI Act in conjunction with the MSR.

5. The Competence Question

165 From a legal and also from a practical perspective it is essential to determine who will be responsible for continuous monitoring, for random testing, and for the penalty system. Generally, the role of a review entity (and in particular the CA introduced in Section B.II.2.a) could either be taken up by a state authority (or authorised third party), the AI deployer, or the user.

159 Regulation of the European Parliament and of the Council 2019/1020 of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, 2019 O.J. (L 169) 1.

166 A state authority (or authorised third party) would have the advantage of being neutral and having no conflict of interest. Furthermore, since we do not suggest checking *all* outputs of *all* AI systems, but instead do random sampling from a limited number of (Big) AI systems, the workforce and expertise required is limited, especially when considering the benefits associated with the monitoring system. The effort required to equip government agencies as well as third parties sufficiently to carry out monitoring would therefore be moderate. In addition to that, state authorities would be the only party who could act as market surveillance authority and have the power to impose fines.

167 AI deployers, on the other hand, generally have a higher level of expertise with regard to the AI systems they have built and run and will tend to be better able to increase the workforce. It is clear who they are, compared to AI providers who may be numerous and difficult to identify in the supply chain. And self-certification is already a process used in other industries. However, shifting the competency to the deployers could lead to a regulatory overload (especially for SMEs). In addition to that, it is questionable how effective this form of self-regulation will be. Finally, the user appears to be the least suitable party, as the user lacks the expertise, resources and competence to impose fines.

168 Hence, imposing a penalty system can and should only be done by the responsible state authorities. Testing and monitoring require more legal than technical knowledge, once the system is installed. For the system to work, it is essential to guarantee independence and impartiality of the testing entity. If the testing is done by AI deployers, the testing entity must have a position comparable to the data protection officer (DPO).

D. Conclusion

169 This paper presents new and alternative avenues for regulating AI, understood in the broad sense of state intervention that directly or indirectly affects how AI is developed, operated and deployed. In doing so, it also identifies two new regulatory axes to be addressed - Big/Small AI and Big/Small Teams - and the challenges associated with designing tools that are effective and tailored to these axes. Against this background, our proposals in this paper thus fall into two broad categories:

- Forms of strict liability, as a means to incentivise operators of Small (and Big) AI systems (where direct regulation is unlikely to be effective, and often unnecessary) to take steps preventing or at least minimising harm resulting from their AI.

- A mechanism of constant monitoring through random sampling with human testing and determination, as well as associated cumulative "penalty points", to control the impact of Big AI systems on individuals and society before they cause large-scale, and in particular systematic, harm to society.

170 Our proposals aim to have the maximal possible impact on the behaviour of AI operators and on where and how AI is used, so as to prevent harm and encourage responsible development and use of AI. In our view, these proposals are implementable, policeable, enforceable, cost-effective, and technologically-neutral; they do not focus on what the AI *is*, but rather what it *does*.

171 While it is the (impact) size of the AI that appears to be the focus here, the size of the Team and the high (but not absolute) correlation between the two help us to understand the risks and challenges of *what* AI is doing, and *who* is doing it. Ultimately, our proposals have been designed to address the risks outlined in our 2x2 matrix of AI types. Whether the regulatory and legal implementation tools we discuss will suffice *on their own* to address our concerns remains to be seen, and indeed we propose these tools to operate alongside or concurrently with others. However, what is clear in our view is that any overarching AI regulatory system which overlooks or discounts the risks outlined in our 2x2 matrix of AI types will almost certainly leave "regulatory gaps" through which AI will undoubtedly slip and cause harm.

172 Turning to the current framework for implementation and recent legislative action on AI, in particular the EU AI Act, the constant monitoring mechanism arguably may be realised as part of the AI Act's provider obligations regarding high-risk AI systems. However, the concept of risk classification under the AI Act does not perfectly coincide with our proposed concept of Big AI/Small AI. The penalty point system may be realised as part of the existing EU market surveillance regime. Concerning private (tort) liability, the old laws only provide for negligence-based responsibility. However, the new Product Liability Directive might even lead to a situation where operators of all types of AI can be held liable to an appropriate extent. We hope our contribution stimulates the debate about regulatory strategies concerning AI and invite other researchers to devise further approaches.

173 With regard to very large online platforms and search engines (which generally will rely on AI and other means of automation for content recommendation and moderation, and in supplying their services more generally), the enhanced responsibility framework under the DSA serves as a useful starting point for implementing our proposals. These platforms could implement an 'internal' version of our monitoring system as part of their risk management commitments - while the auditing and other external controls foreseen by the DSA fall significantly short of the constant monitoring with human oversight our check authority is meant to undertake. At the same time, the potential for broader and often diffuse societal harm posed by large platforms and their reliance on AI will often not be detectable and hence preventable through a constant monitoring system that necessarily produces swift results based on tests for harms primarily observable from the AI output alone. In light of what is at stake for individuals and societies, we certainly hope that all actors involved seriously engage with the risk assessment and auditing tools envisioned by the DSA.[160]

174 And finally, the risk to society of AI-related technologies is now well-appreciated by many. However, the focus of such awareness is often on a single, high-impact AI or high-impact event, such as AI starting a (nuclear) war, AI taking over (all) digital systems, an AI-powered robot army rising up, or similar dystopian scenarios. Our approach deals with the AI risk posed by a "death by 1000 cuts", whereby either a vast number of (undetectable) AI systems (Small AI) start creating overall weaknesses in society through their undesirable and harmful impact, or one AI system does likewise through a vast series of mini-impacts (Big AI), no single one of which is catastrophic but in aggregate can systematically shift how society operates, potentially for the worse. Each of these pose the same sort of threat to humanity that mosquitoes have throughout history; spreading harm to many, and difficult to detect and deal with individually. No one mosquito can be considered a significant threat, but it is estimated that over half of human deaths through history can be attributed to mosquitoes.[161] Without rapid, decisive, and impactful action, humanity may find itself facing the same sort of "undetectable" threat that it faced in the past with mosquitoes: people around us will start becoming ill, perhaps even dying and we won't really know why, or how to stop it.

160 Early analysis of DSA compliance by VLOPSEs suggests that this not an easy task in practice; see Sekwenz et al., *From Reports to Reality: Testing Consistency in Instagram's Digital Services Act Compliance Data*, 2025, arXiv:2507.01787.

161 Winegard, *The Mosquito: A Human History of Our Deadliest Predator*, E P Dutton (Aug. 6, 2019).